Fall 1998

Dr. Aron Katsenelinboigen and Dr. Vera Zubarev

Syllabus of the course

The Art of Decision Making

A. THEORY.

1. A. Katsenelinboigen, Indeterministic Economics. New York: Praeger Publ., 1992.
2. A. Katsenelinboigen, The Concept of Indeterminism. New York: Praeger Publ., 1997.
3. An Aesthteic Approach to..., ed. by A. Katsenelinboigen and M. Mandelbaum. Philadelphia: Frends of Art, 1997.

Session 1. Introduction to the course. Typology of the methods of decision making.
1. P. Kleindorfer, H. Kunreuther, P. Schoemaker, Decision sciences: an integrative perspective. New York : Cambridge University Press, 1993.
2. Simon, H., Models of Bounded Rationality, Cambridge, MIT Press, 1982, Vol 2.

Session 2-3. Chess as model for a multi-stage decision making process
1. Katsenelinboigen, A., "Chess as a Multi-Stage Game." Selected Topics in Indeterministic Systems, Seaside, CA: Intersystems Publ,1989.
2. Shannon, C., "Programming a Computer for Playing Chess." Philosophical Magazine, No. 41, 1950.

Session 4-6. Decision making on the corporate level

a) Positioning a corporation
1. Armstong, S. and Collopy, F. "Competitor Orientation: Effects of Objectives and Information on Managerial Decisions and Profitability." Journal of Marketing Research, 1996, Vol. XXXIII.
2. Baumol, W., Business Behavior, Value and Growth, New York: Marcourt, 1967.
3. Leibenstein, H., "A Branch of Economics is Missing: Micro-Micro Theory." Journal of Economic Literature, June 1979, Vol. XV.
4. Schoeffler,S. "The Role of Science in Business Strategy." Handbook of Business Strategy. Boston: Warren, Gorham, and Lamont, 1991.
5. Business Chess, Perspectives, The Boston Consulting Group, 1984.
6. Miniter, M. "The Myth of Market Share." The Wall Street Journal

b) Combinational style
1. Drucker, P. Innovation and Entrepreneurship, New York: Harper & Row, 1985.

c) The concept of interactive planning. Directing, potentiating and optimizing
1. Ackoff, R. "A Typology of Planning", Creating the Corporate Future, New York: John Wiley, 1981.
2. Mintzberg, H., "The Fall and Rise of Strategic Planning", Harvard Business Review, January, 1994

B. APPLICATIONS TO LITERATURE

Session 7. Introductory. Text, context, subtext. Randomness and Fate.
Session 8. *Romeo and Juliet* by Shakespeare.
Session 9. Discussions.
Session 10. *Romeo and Juliet* by Zeffirelly. Movie and discussions.
Session 11. *Romeo and Juliet* by Luhrman. Movie and discussion.
Session 12. Predispositioning in art and other systems.

C. PRESENTATIONS **by the students based on midterm and final papers,**

Session 13-14.

1847

1

WHARTON REPROGRAPHICS

CREATING THE CORPORATE FUTURE

PLAN OR BE
PLANNED FOR

RUSSELL L. ACKOFF

The Wharton School
University of Pennsylvania

JOHN WILEY & SONS
New York · Chichester · Brisbane · Toronto

Recall that inactivists are willing to settle for doing well enough, to *satisfice*; preactivists want to do as well as possible, to *optimize*. Interactivists want to do better in the future than the best we are capable of doing now, to *idealize*. Therefore, interactivists focus on improving performance over time rather than on how well they can do at a particular time under particular conditions. Their objective is to maximize their ability to learn and adapt, *to develop*.

Learning and adaptation are taken as key requirements because of the accelerating rate of social change. No societal or corporate problem stays solved for long, and the duration of the solution reduces as the rate of change increases. Furthermore, the solution to any problem creates new and often more difficult problems. The progress of science, for example, has derived as much from the progression from simple to complex problems as it has from the progression from complex to simple solutions.

Recall that inactivists focus on not doing things that should not be done (avoiding errors of commission), and preactivists on doing things that should be done (avoiding errors of omission). Inactivists try to avoid both but they are more concerned with two other types of error that are generally ignored: asking the wrong questions or solving the wrong problems and not asking or not solving the right ones. They believe we fail more often because of an inability to face the right problems than because of an inability to solve the problems we face.

The interactivist believes our failure to address the right problems arises from a lack of awareness of what we are striving for *ultimately*. Humans are more than ends-seeking animals; we are *ideal-seeking*. Curiously, however, this characteristic of humans is ignored in all approaches to planning other than the interactive.

There are three types of ends that people pursue:

1. *Goals*: those ends that we can expect to attain within the period covered by planning.
2. *Objectives*: those ends that we do not expect to attain within the period planned for but which we hope to attain later, and toward which we believe progress is possible within the period planned for.
3. *Ideals*: those ends that are believed to be unattainable but towards which we believe progress is possible during and after the period planned for.

Planning ought to involve explicitly all three types of ends, but it seldom does. Which types are taken into account determine whether planning is *operational, tactical, strategic,* or *normative* (see Table 3.2).

Table 3.2 Types of Planning and Planning Postures

Types of planning	Means	Goals	Objectives	Ideals	Associated with
Operational	Choose	Given	Given	Given	Inactivism
Tactical	Choose	Choose	Given	Given	Reactivism
Strategic	Choose	Choose	Choose	Given	Preactivism
Normative	Choose	Choose	Choose	Choose	Interactivism

Operational planning consists of selecting means for pursuing goals that are either given, set, or imposed by a higher authority, or are accepted by convention; for example, planning to produce an amount of product specified by a higher authority. Such planning is usually short-range.

Inactivists engage in operational planning despite their negative attitude toward planning in general. Their goal, to keep things as they are, is given, but they must select the means by which this is to be accomplished.

Tactical planning consists of selecting means and goals for pursuing objectives that are either given, set, or imposed by a higher authority, or are accepted by convention. For example, a company's objective, to obtain market leadership in the next ten years, may be imposed on its marketing department. That department may then decide to reduce the gap between the company and the current market leader by a specified amount (its goal) in the period covered by its five-year plan. It then selects the means by which it will try to do so. Such planning tends to be medium-range.

Reactivists engage in tactical planning. They must select the previous states to which to return (their goals) and, of course, the means to do so.

Strategic planning consists of selecting means, goals, and objectives, but ideals are either given, imposed by a higher authority, accepted by convention, or, as is usually the case, not formulated. Such planning tends to be long-range.

Preactivists engage in strategic planning, holding in view a period longer than the one for which planning is done. In this sense they are more far-sighted than either reactivists or inactivists.

Normative planning requires the explicit selection of means, goals, objectives, and ideals. Such planning is indefinitely extended. It has no fixed horizon.

Interactivists engage in normative planning. As will be seen, not only do ideals play an important role in such planning, but they play the *key* role.

2

WHARTON REPROGRAPHICS

Planners shouldn't create strategies, but they can supply data, help managers think strategically, and program the vision.

The Fall and Rise of Strategic Planning

by Henry Mintzberg

When strategic planning arrived on the scene in the mid-1960s, corporate leaders embraced it as "the one best way" to devise and implement strategies that would enhance the competitiveness of each business unit. True to the scientific management pioneered by Frederick Taylor, this one best way involved separating thinking from doing and creating a new function staffed by specialists: strategic planners. Planning systems were expected to produce the best strategies as well as step-by-step instructions for carrying out those strategies so that the doers, the managers of businesses, could not get them wrong. As we now know, planning has not exactly worked out that way.

Strategic planning isn't strategic thinking. One is analysis, and the other is synthesis.

While certainly not dead, strategic planning has long since fallen from its pedestal. But even now, few people fully understand the reason: *strategic planning* is not *strategic thinking*. Indeed, strategic planning often spoils strategic thinking, causing managers to confuse real vision with the manipulation of numbers. And this confusion lies at the heart of the issue: the most successful strategies are visions, not plans.

Strategic planning, as it has been practiced, has really been *strategic programming*, the articulation and elaboration of strategies, or visions, that already exist. When companies understand the difference between planning and strategic thinking, they can get back to what the strategy-making process should be: capturing what the manager learns from all sources (both the soft insights from his or her personal experiences and the experiences of others throughout the organization and the hard data from market research and the like) and then synthesizing that learning into a vision of the direction that the business should pursue.

Henry Mintzberg is professor of management at McGill University in Montreal, Quebec, and visiting professor at INSEAD in Fontainebleau, France. This article, his fifth contribution to HBR, is adapted from his latest book, The Rise and Fall of Strategic Planning *(Free Press and Prentice Hall International, 1994).*

Copyright © 1993 by the President and Fellows of Harvard College. All rights reserved.

Organizations disenchanted with strategic planning should not get rid of their planners or conclude that there is no need for programming. Rather, organizations should transform the conventional planning job. Planners should make their contribution *around* the strategy-making process rather than *inside* it. They should supply the formal analyses or hard data that strategic thinking requires, as long as they do it to broaden the consideration of issues rather than to discover the one right answer. They should act as catalysts who support strategy making by aiding and encouraging managers to think strategically. And, finally, they can be programmers of a strategy, helping to specify the series of concrete steps needed to carry out the vision.

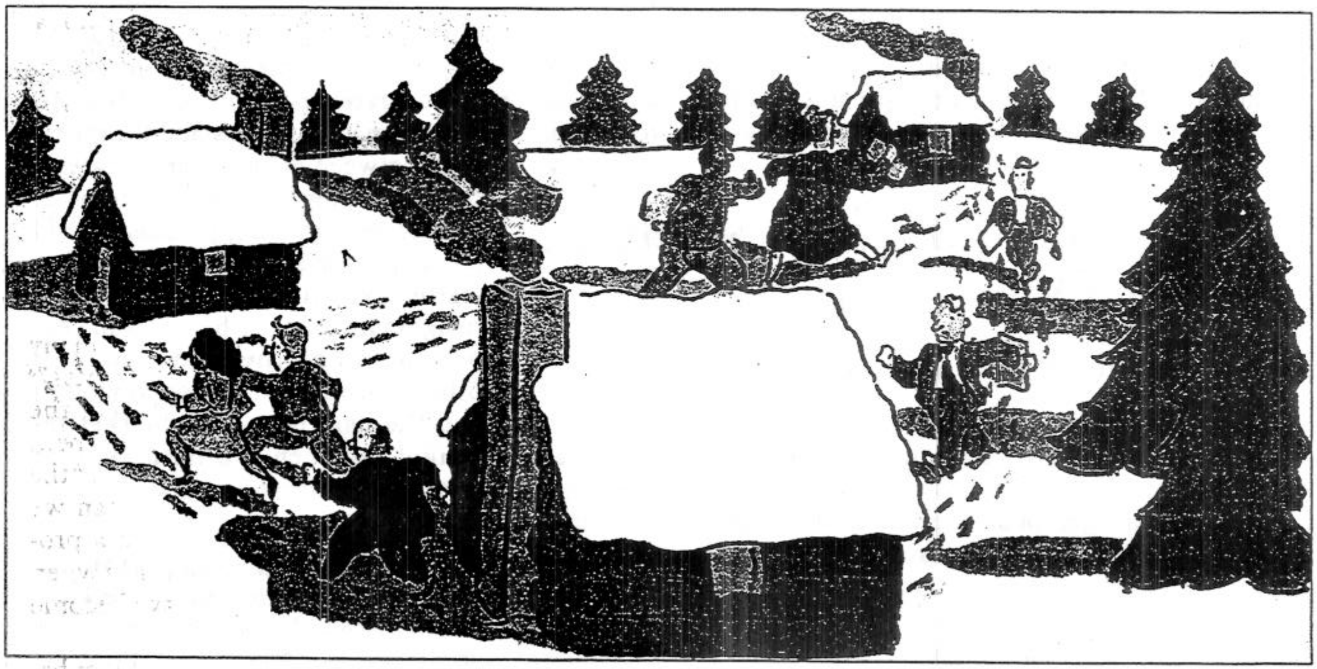
Planners should make their greatest contribution *around* the strategy-making process rather than *inside* it.

By redefining the planner's job, companies will acknowledge the difference between planning and strategic thinking. Planning has always been about *analysis* – about breaking down a goal or set of intentions into steps, formalizing those steps so that they can be implemented almost automatically, and articulating the anticipated consequences or results of each step. "I favour a set of analytical techniques for developing strategy," Michael Porter, probably the most widely read writer on strategy, wrote in the *Economist*.[1]

The label "strategic planning" has been applied to all kinds of activities, such as going off to an informal retreat in the mountains to talk about strategy. But call that activity "planning," let conventional planners organize it, and watch how quickly the event becomes formalized (mission statements in the morning, assessment of corporate strengths and weaknesses in the afternoon, strategies carefully articulated by 5 P.M.).

Strategic thinking, in contrast, is about *synthesis.* It involves intuition and creativity. The outcome of strategic thinking is an integrated perspective of the enterprise, a not-too-precisely articulated vision of direction, such as the vision of Jim Clark, the founder of Silicon Graphics, that three-dimensional visual computing is the way to make computers easier to use.

Such strategies often cannot be developed on schedule and immaculately conceived. They must be free to appear at any time and at any place in the organization, typically through messy processes of informal learning that must necessarily be carried out by people at various levels who are deeply involved with the specific issues at hand.



Call an informal retreat "planning," let conventional planners organize it, and watch how quickly the event becomes formalized.

DRAWINGS BY GARISON WEILAND

Formal planning, by its very analytical nature, has been and always will be dependent on the preservation and rearrangement of established categories – the existing levels of strategy (corporate, business, functional), the established types of products (defined as "strategic business units"), overlaid on the current units of structure (divisions, departments, etc.). But real strategic change requires not merely rearranging the established categories, but inventing new ones.

Search all those strategic planning diagrams, all those interconnected boxes that supposedly give you strategies, and nowhere will you find a single one that explains the creative act of synthesizing experiences into a novel strategy. Take the example of the Polaroid camera. One day in 1943, Edwin Land's three-year-old daughter asked why she could not immediately see the picture he had just taken of her. Within an hour, this scientist conceived the camera that would transform his company. In other words, Land's vision was the synthesis of the insight evoked by his daughter's question and his vast technical knowledge.

Real strategic change requires inventing new categories, not rearranging old ones.

Strategy making needs to function beyond the boxes, to encourage the informal learning that produces new perspectives and new combinations. As the saying goes, life is larger than our categories. Planning's failure to transcend the categories explains why it has discouraged serious organizational change. This failure is why formal planning has promoted strategies that are extrapolated from the past or copied from others. Strategic planning has not only never amounted to strategic thinking but has, in fact, often impeded it. Once managers understand this, they can avoid other costly misadventures caused by applying formal technique, without judgment and intuition, to problem solving.

The goal of those who promote planning is to reduce managers' power over strategy making.

The Pitfalls of Planning

If you ask conventional planners what went wrong, they will inevitably point to a series of pitfalls for which they, of course, are not responsible. Planners would have people believe that planning fails when it does not receive the support it deserves from top management or when it encounters resistance to change in the organization. But surely no technique ever received more top management support than strategic planning did in its heyday. Strategic planning itself has discouraged the commitment of top managers and has tended to create the very climates its proponents have found so uncongenial to its practice.

The problem is that planning represents a *calculating* style of management, not a *committing* style. Managers with a committing style engage people in a journey. They lead in such a way that everyone on the journey helps shape its course. As a result, enthusiasm inevitably builds along the way. Those with a calculating style fix on a destination and calculate what the group must do to get there, with no concern for the members' preferences. But calculated strategies have no value in and of themselves; to paraphrase the words of sociologist Philip Selznick, strategies take on value only as committed people infuse them with energy.[2]

No matter how much lip service has been paid to the contrary, the very purpose of those who promote conventional strategic planning is to reduce the power of management over strategy making. George Steiner declared, "If an organization is managed by intuitive geniuses there is no need for formal strategic planning. But how many organizations are so blessed? And, if they are, how many times are intuitives correct in their judgments?"[3] Peter Lorange, who is equally prominent in the field, stated, "The CEO should typically not be... deeply involved" in the process, but rather be "the designer of [it] in a general sense."[4] How can we expect top managers to be committed to a process that depicts them in this way, especially when its failures to deliver on its promises have become so evident?

At lower levels in the hierarchy, the problem becomes more severe because planning has often been used to exercise blatant control over business managers. No wonder so many middle managers have welcomed the overthrow of strategic planning. All they wanted was a commitment to their own business strategies without having to fight the planners to get it!

The Fallacies of Strategic Planning

An expert has been defined as someone who avoids the many pitfalls on his or her way to the grand fallacy. For strategic planning, the grand fallacy is this: because analysis encompasses synthesis, strategic planning is strategy making. This fallacy itself rests on three fallacious assumptions: that prediction is possible, that strategists can be detached from the subjects of their strategies, and, above all, that the strategy-making process can be formalized.

The Fallacy of Prediction. According to the premises of strategic planning, the world is supposed to hold still while a plan is being developed and then stay on the predicted course while that plan is being implemented. How else to explain those lockstep schedules that have strategies appearing on the first of June, to be approved by the board of directors on the fifteenth? One can just picture competitors waiting for the board's approval, especially if they are Japanese and don't believe in such planning to begin with.

In 1965, Igor Ansoff wrote in his influential book *Corporate Strategy*, "We shall refer to the period for which the firm is able to construct forecasts with an accuracy of, say, plus or minus 20 percent as the *planning horizon* of the firm."[5] What an extraordinary statement! How in the world can any company know the period for which it can forecast with a given accuracy?

The evidence, in fact, points to the contrary. While certain repetitive patterns, such as seasons, may be predictable, the forecasting of discontinuities, such as a technological innovation or a price increase, is virtually impossible. Of course, some people sometimes "see" such things coming. That is why we call them "visionaries." But they create their strategies in much more personalized and intuitive ways.

The Fallacy of Detachment. In her book *Institutionalizing Innovation*, Mariann Jelinek developed the interesting point that strategic planning is to the executive suite what Taylor's work-study methods were to the factory floor – a way to circumvent human idiosyncrasies in order to systematize behavior. "It is through administrative systems that planning and policy are made possible, because the systems capture knowledge *about* the task." Thus "true management by exception, and true policy direction are now possible, solely because management is no longer wholly immersed in the details of the task itself."[6]

According to this viewpoint, if the system does the thinking, then strategies must be detached from operations (or "tactics"), formulation from implementation, thinkers from doers, and so strategists from the objects of their strategies.

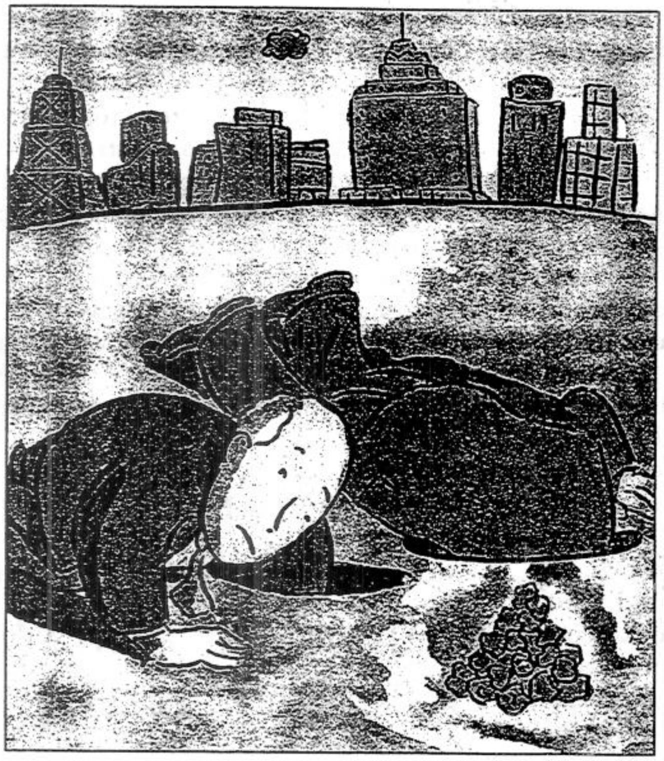
The trick, of course, is to get the relevant information up there, so that senior managers on high can be informed about the details down below without having to immerse themselves in them. Planners' favored solution has been "hard data," quantitative aggregates of the detailed "facts" about the organization and its context, neatly packaged and regularly delivered. With such information, senior managers need never leave their executive suites or planners their staff offices. Together they can formulate – work with their heads – so that the hands can get on with implementation.

All of this is dangerously fallacious. Innovation has never been institutionalized. Systems have never been able to reproduce the synthesis created by the genius entrepreneur or even the ordinary competent strategist, and they likely never will.

Where in the planning literature is there a shred of evidence that anyone has bothered to find out how managers make strategies?

Ironically, strategic planning has missed one of Taylor's most important messages: work processes must be fully understood before they can be formally programmed. But where in the planning literature is there a shred of evidence that anyone has ever bothered to find out how it is that managers really do make strategies? Instead many practitioners and theorists have wrongly assumed that strategic planning, strategic thinking, and strategy making are all synonymous, at least in best practice.

The problem with the hard data that are supposed to inform the senior manager is they can have a decidedly soft underbelly. Such data take time to harden, which often makes them late. They tend to



Real strategists get their hands dirty digging for ideas, and real strategies are built from the nuggets they uncover.

lack richness; for example, they often exclude the qualitative. And they tend to be overly aggregated, missing important nuances. These are the reasons managers who rely on formalized information, such as market-research reports or accounting statements in business and opinion polls in government, tend to be detached in more ways than one. Study after study has shown that the most effective managers rely on some of the softest forms of information, including gossip, hearsay, and various other intangible scraps of information.

My research and that of many others demonstrates that strategy making is an immensely complex process, which involves the most sophisticated, subtle, and, at times, subconscious elements of human thinking.

A strategy can be deliberate. It can realize the specific intentions of senior management, for example, to attack and conquer a new market. But a strategy can also be emergent, meaning that a convergent pattern has formed among the different actions taken by the organization one at a time.

In other words, strategies can develop inadvertently, without the conscious intention of senior management, often through a process of learning. A salesperson convinces a different kind of customer to try a product. Other salespeople follow up with their customers, and the next thing management knows, its products have penetrated a new market. When it takes the form of fits and starts, discoveries based on serendipitous events, and the recognition of unexpected patterns, learning inevitably plays *a,* if not *the,* crucial role in the development of novel strategies.

Contrary to what traditional planning would have us believe, deliberate strategies are not necessarily good, nor are emergent strategies necessarily bad. I believe that all viable strategies have emergent and deliberate qualities, since all must combine some degree of flexible learning with some degree of cerebral control.

Vision is unavailable to those who cannot "see" with their own eyes. Real strategists get their hands dirty digging for ideas, and real strategies are built from the occasional nuggets they uncover. These are not people who abstract themselves from the daily details; they are the ones who immerse themselves in them while being able to abstract the strategic messages from them. The big picture is painted with little strokes.

The Fallacy of Formalization. The failure of strategic planning is the failure of systems to do better than, or even nearly as well as, human beings. Formal systems, mechanical or otherwise, have offered no improved means of dealing with the information overload of human brains; indeed, they have often made matters worse. All the promises about artificial intelligence, expert systems, and the like improving if not replacing human intuition never materialized at the strategy level. Formal systems could certainly process more information, at least hard information. But they could never *internalize* it, *comprehend* it, *synthesize* it. In a literal sense, planning could not learn.

Formalization implies a rational sequence, from analysis through administrative procedure to eventual action. But strategy making as a learning process can proceed in the other direction too. We think in order to act, to be sure, but we also act in order to think. We try things, and those experiments that work converge gradually into viable patterns that become strategies. This is the very essence of strategy making as a learning process.

Formal procedures will never be able to forecast discontinuities, inform detached managers, or create novel strategies. Far from providing strategies, planning could not proceed without their prior existence. All this time, therefore, strategic planning has been misnamed. It should have been called strategic programming, distinguished from other useful things that planners can do, and promoted as a process to formalize, when necessary, the consequences of strategies that have already been developed. In short, we should drop the label "strategic planning" altogether.

Planning, Plans, and Planners

Two important messages have been conveyed through all the difficulties encountered by strategic planning. But only one of them has been widely accepted in the planning community: business-unit managers must take full and effective charge of the strategy-making process. The lesson that has still not been accepted is that managers will never be able to take charge through a formalized process. What then can be the roles for planning, for plans, and for planners in organizations?

> Sometimes strategies must be left as broad visions, not precisely articulated, to adapt to a changing environment.

Planners and managers have different advantages. Planners lack managers' authority to make commitments, and, more important, managers' access to soft information critical to strategy making. But because of their time pressures, managers tend to favor action over reflection and the oral over the written, which can cause them to overlook important analytical information. Strategies cannot be created by analysis, but their development can be helped by it.

Planners, on the other hand, have the time and, most important, the inclination to analyze. They have critical roles to play alongside line managers, but not as conventionally conceived. They should work in the spirit of what I like to call a "soft analyst," whose intent is to pose the right questions rather than to find the right answers. That way, complex issues get opened up to thoughtful consideration instead of being closed down prematurely by snap decisions.

Planning as Strategic Programming. Planning cannot generate strategies. But given viable strategies, it can program them; it can make them operational. For one supermarket chain that a colleague and I studied, planning was the articulation, justification, and elaboration of the strategic vision that the company's leader already had. Planning was not deciding to expand into shopping centers, but explicating to what extent and when, with how many stores, and on what schedule.

An appropriate image for the planner might be that person left behind in a meeting, together with the chief executive, after everyone else has departed. All of the strategic decisions that were made are symbolically strewn about the table. The CEO turns to the planner and says, "There they all are; clean them up. Package them neatly so that we can tell everyone about them and get things going." In more formal language, strategic programming involves three steps: codification, elaboration, and conversion of strategies.

Codification means clarifying and expressing the strategies in terms sufficiently clear to render them formally operational, so that their consequences can be worked out in detail. This requires a good deal of interpretation and careful attention to what might be lost in articulation: nuance, subtlety, qualification. A broad vision, like capturing the market for a new technology, is one thing, but a specific plan–35% market share, focusing on the high end–is quite another.

Elaboration means breaking down the codified strategies into substrategies and ad hoc programs as well as overall action plans specifying what must be done to realize each strategy: build four new factories and hire 200 new workers, for example.

And conversion means considering the effects of the changes on the organization's operations–effects on budgets and performance controls, for example. Here a kind of great divide must be crossed from the nonroutine world of strategies and programs to the routine world of budgets and objectives. Objectives have to be restated and budgets reworked, and policies and standard operating procedures reconsidered, to take into account the consequences of the specific changes.

One point must be emphasized. Strategic programming is not "the one best way" or even necessarily a good way. Managers don't always need to program their strategies formally. Sometimes they must leave their strategies flexible, as broad visions, to adapt to a changing environment. Only when an organization is sure of the relative stability of its environment and is in need of the tight coordination of a myriad of intricate operations (as is typically the case of airlines with their needs for complicated scheduling), does such strategic programming make sense.

Plans as Tools to Communicate and Control. Why program strategy? The most obvious reason is for coordination, to ensure that everyone in the organization pulls in the same direction. Plans in the form of programs – schedules, budgets, and so on – can be prime media to communicate strategic intentions and to control the individual pursuit of them, in so far, of course, as common direction is considered to be more important than individual discretion.

Plans can also be used to gain the tangible as well as moral support of influential outsiders. Written plans inform financiers, suppliers, government agencies, and others about the intentions of the organization so that these groups can help it achieve its plans.

Planners as Strategy Finders. As noted, some of the most important strategies in organizations emerge without the intention or sometimes even the awareness of top managers. Fully exploiting these strategies, though, often requires that they be recognized and then broadened in their impact, like taking a new use for a product accidentally discovered by a salesperson and turning it into a major new business. It is obviously the responsibility of managers to discover and anoint these strategies. But planners can assist managers in finding these fledgling strategies in their organizations' activities or in those of competing organizations.

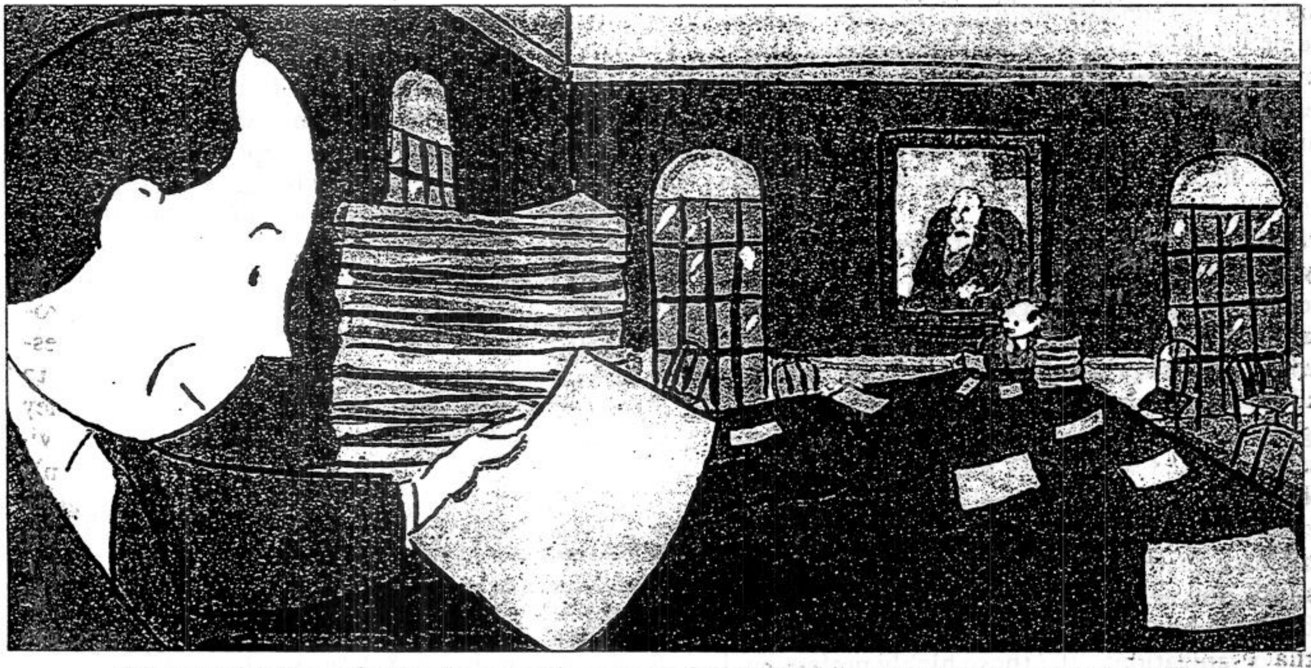
Planners can snoop around places they might not normally visit to find patterns amid the noise of failed experiments, seemingly random activities, and messy learning. They can discover new ways of doing or perceiving things, for example, spotting newly uncovered markets and understanding their implied new products.

Planners as Analysts. In-depth examinations of what planners actually do suggests that the effective ones spend a good deal of time not so much doing or even encouraging planning as carrying out analyses of specific issues. Planners are obvious candidates for the job of studying the hard data and ensuring that managers consider the results in the strategy-making process.

Much of this analysis will necessarily be quick and dirty, that is, in the time frame and on the ad hoc basis required by managers. It may include industry or competitive analyses as well as internal studies, including the use of computer models to analyze trends in the organization.

But some of the best models that planners can offer managers are simply alternative conceptual interpretations of their world, such as a new way to view the organization's distribution system. As Arie de Geus, the one-time head of planning at Royal Dutch/Shell, wrote in his HBR article "Planning as Learning" (March-April 1988), "The real purpose of effective planning is not to make plans but to change the...mental models that...decision makers carry in their heads."

Planners as Catalysts. The planning literature has long promoted the role of catalyst for the planner, but not as I will describe it here. It is not planning that planners should be urging on their organizations so much as any form of behavior that can lead to effective performance in a given situation.



An appropriate image for the planner is the person left behind in a meeting, together with the CEO, who helps package the strategic decisions that are symbolically strewn about the table.

Sometimes that may even mean criticizing formal planning itself.

When they act as catalysts, planners do not enter the black box of strategy making; they ensure that the box is occupied with active line managers. In other words, they encourage managers to think about the future in creative ways.

Such planners see their job as getting others to question conventional wisdom and especially helping people out of conceptual ruts (which managers with long experience in stable strategies are apt to dig themselves into). To do their jobs, they may have to use provocation or shock tactics like raising difficult questions and challenging conventional assumptions.

Left- and Right-Handed Planners

Two very different kinds of people populate the planning function. One is an analytic thinker, who is closer to the conventional image of the planner. He or she is dedicated to bringing order to the organization. Above all, this person programs intended strategies and sees to it that they are communicated clearly. He or she also carries out analytic studies to ensure consideration of the necessary hard data and carefully scrutinizes strategies intended for implementation. We might label him or her the *right-handed planner.*

The second is less conventional but present nonetheless in many organizations. This planner is a creative thinker who seeks to open up the strategy-making process. As a "soft analyst," this planner is prepared to conduct more quick and dirty studies. He or she likes to find strategies in strange places and to encourage others to think strategically. This person is somewhat more inclined toward the intuitive processes identified with the brain's right hemisphere. We might call him or her the *left-handed planner.*

Many organizations need both types, and it is top management's job to ensure that it has them in appropriate proportions. Organizations need people to bring order to the messy world of management as well as challenge the conventions that managers and especially their organizations develop. Some organizations (those big, machine-like bureaucracies concerned with mass production) may favor the right-handed planners, while others (the loose, flexible "adhocracies," or project organizations) may favor the left-handed ones. But both kinds of organization need both types of planners, if only to offset their natural tendencies. And, of course, some organizations, like those highly professionalized hospitals and educational systems that have been forced to waste so much time doing ill-conceived strategic planning, may prefer to have very few of either!

The Formalization Edge

We human beings seem predisposed to formalize our behavior. But we must be careful not to go over the formalization edge. No doubt we must formalize to do many of the things we wish to in modern society. That is why we have organizations. But the experiences of what has been labeled strategic planning teach us that there are limits. These limits must be understood, especially for complex and creative activities like strategy making.

Strategy making is not an isolated process. It does not happen just because a meeting is held with that label. To the contrary, strategy making is a process interwoven with all that it takes to manage an organization. Systems do not think, and when they are used for more than the facilitation of human thinking, they can prevent thinking.

Three decades of experience with strategic planning have taught us about the need to loosen up the process of strategy making rather than trying to seal it off by arbitrary formalization. Through all the false starts and excessive rhetoric, we have learned what planning is not and what it cannot do. But we have also learned what planning is and what it can do, and perhaps of greater use, what planners themselves can do beyond planning. We have also learned how the literature of management can get carried away and, more important, about the appropriate place for analysis in organizations.

The story of strategic planning, in other words, has taught us not only about formal technique itself but also about how organizations function and how managers do and don't cope with that functioning. Most significant, it has told us something about how we think as human beings, and that we sometimes stop thinking.

References

1. Michael Porter, "The State of Strategic Thinking," *Economist*, May 23, 1987, p. 21.

2. Philip Selznick, *Leadership in Administration: A Sociological Interpretation* (New York: Harper & Row, 1957).

3. George Steiner, *Strategic Planning: What Every Manager* Must *Know* (New York: Free Press, 1979), p. 9.

4. Peter Lorange, "Roles of the CEO in Strategic Planning and Control Processes," in a seminar on The Role of General Management in Strategy Formulation and Evaluation, cosponsored by E.S.S.E.C., E.I.A.S.M., and I.A.E. (Cergy, France: April 28-30, 1980), p. 2.

5. H. Igor Ansoff, *Corporate Strategy: An Analytic Approach to Business Policy for Growth and Expansion* (New York: McGraw-Hill, 1965), p. 44.

6. Mariann Jelinek, *Institutionalizing Innovation: A Study of Organizational Learning Systems* (New York: Praeger, 1979), p. 139.

Reprint 94107

Indeterministic Economics

Aron Katsenelinboigen

PRAEGER
New York
Westport, Connecticut
London

WHARTON REPROGRAPHICS

Library of Congress Cataloging-in-Publication Data

Katsenelinboigen, Aron.
Indeterministic economics / Aron Katsenelinboigen.
p. cm.
Includes bibliographical references and index.
ISBN 0-275-94143-4 (alk. paper)
1. Economics—Methodology. I. Title.
HB131.K36 1992
330–dc20 91-27722

British Library Cataloguing in Publication Data is available.

Copyright © 1992 by Aron Katsenelinboigen

All rights reserved. No portion of this book may be
reproduced, by any process or technique, without the
express written consent of the publisher.

Library of Congress Catalog Card Number: 91-27722
ISBN: 0-275-94143-4

First published in 1992

Praeger Publishers, One Madison Avenue, New York, NY 10010
An imprint of Greenwood Publishing Group, Inc.

Printed in the United States of America

The paper used in this book complies with the
Permanent Paper Standard issued by the National
Information Standards Organization (Z39.48–1984).

10 9 8 7 6 5 4 3 2 1

To Anatol Rapoport —

a great man and dear friend

on his 80th birthday

Creative Potential

Beyond
the omega
Strange little letters lie
In black hidey-holes
Waiting
For universes
Yet unmade,
uncreated,
Lacking the word

All in all
They have much
Alphabetical potential
But the problem
(as
always)

Is
how to string
The little letters
Together
And where to leave
The equally vital
gaps
The little
blank
Spaces

— A. V. Riasanovsky
Conversations with a Hindu
Philadelphia, 1981. (private distribution)

Contents

Tables and Figures ix
Acknowledgments xi

Introduction 1

PART I — ECONOMIC SYSTEM AND INDETERMINISM

1 Economy as a Dynamic System 19

2 Indeterminism-Determinism 37

3 Degree of Indeterminism and the Performance of an Economic System 56

PART II — DIRECTING AND PROGRAMMING

4 Directing: On the Variety of Ideals and Ideologies 83

5 Programming I: Representing an Economic System 92

6 Programming II: Interpreting the Variety of Modes for Economic System Representation 108

7 Programming III: Methods of Analysis of Economic Systems 125

PART III — POTENTIATING

8 Generalities 137

9 Macroeconomy I: Measuring Economic Potential 166

10 Macroeconomy II: Measuring Economic Development in Kuznets's *Modern Economic Growth* 183

11 Macroeconomy III: Convertibility of an Economic System 192

12 Macroeconomy IV: Mechanisms of Operation of Economic Systems 209

13 Branch: Evaluation of the Potential in Strategic Mineral Resources 221

14 Firm I: Potentiating at the Firm Level 234

15 Firm II: General Quantitative Methods of Evaluation of a Position 249

16 Firm III: Positional and Combinational Styles of Management 263

17 Firm IV: Mechanisms Implementing the Potentiating Method 280

Conclusion 297

Bibliography 303
Author Index 309
Subject Index 311

Tables and Figures

TABLES

I.1 Production Growth Index 9
2.1 Relation between Various Aspects of Systems Approach and the Problematique of Indeterminism 39
3.1 Methods of Performance and Types of Planning 68
8.1 Primary Relational Parameters 140
12.1 Stages of R&D and Manufacturing 209
13.1 Vulnerability Scores 227
13.2 Weights, Scores, and Trends for Aluminum, Cobalt, and Titanium in the Soviet Union and the United States 231
17.1 Correlation between Objectivity and Limitations 293
17.2 Typology of Corporate Leadership 294
C.1 Comparative Evaluation of Parameters Reflecting the Performance of U.S. and Japanese Firms 301

FIGURES

2.1 Matrix of Interrelations of Degree of Indeterminacy and a System's Structures 53
3.1 Matrix of Combinations in the Aesthetic Method 74
4.1 Matrix Representation of Personality Types 87
7.1 Combinations of Certainty and Uncertainty in Formulation and Solution of a Problem 130
8.1 Initial Position 155
8.2 The Tempting Move 155
8.3 The Winning Move 156

Acknowledgments

After immigrating at the end of 1973 to the United States, I immediately got in touch with Anatol Rapoport, the famous systems research professor at the University of Toronto, whom I had met in the USSR. I sent him a transcript (in Russian) of my paper, "Formation of System's Potential," which was the beginning of the present book. I managed to bring this paper out of the USSR through legal but circuitous means. It did not contain any anti-Soviet material. It was a strictly scientific paper, but that still could not guarantee its free passage through Soviet customs. My fears that it may be delayed were justified: a large part of my strictly scientific writing was expunged.

As any provincial scientist to a large degree separated from world science (as I viewed myself in the USSR), I expected Professor Rapoport, one of the famous founders of General Systems Theory, to accuse me of "discovering America" or, at least, to criticize rather substantially my poor knowledge of world science. Instead of his criticism, I received an English translation, by Anatol, prepared for publication in the yearbook *General Systems*, which he edited at the time. I am grateful to Professor Rapoport not only for the publication of this article (as one of the Jewish Passover songs goes, however, in reference to God, any one of his many deeds would have sufficed for gratitude), but also for his attention to my research and my life throughout all these years. I am also grateful to his wife Gwen for her hospitality and tolerance for my extended conversations with Anatol (in Russian, which he speaks fluently, to boot) during my almost annual pilgrimages to their house in Toronto.

I am also grateful to Anatol for introducing me during his 1975 trip to Philadelphia to the great authority on system research, Russell Ackoff, professor at the University of Pennsylvania. I sent Russell my article from the above-mentioned yearbook and received a letter consisting of one sentence, "This is great!" In subsequent years Russell made a large effort to create favorable conditions for my work on this book.

Valuing highly the help of my learned colleagues, I cannot but stress the special role played in my research by my closest friend Norman Gross — one of the pioneers in the area of corporate planning in the United States — and to thank him for the pains he suffered during the years of knowing me.

I would also like to thank the National Council for Soviet and East European Research, for their grant, given to me in 1980–1982 for the development of a methodology of assessing Soviet economic potential, and Professor Herbert Levine, for his help in preparing grant applications and in editing the final report.

I would like to thank Dr. Abraham Becker for twice giving me the opportunity to express my views on the theory of development of economic potential before the employees of the RAND Corporation. Recognizing the necessity of preparation for a presentation in such a prestigious institution is sufficient to realize how stimulating these opportunities were for my work on the concept of potential. My conversations with Drs. Becker and Nancy Nimitz were also quite helpful.

I am grateful to Dr. Nancy Bauer, the former director of the Graduate Program of Dynamics of Organization at the University of Pennsylvania, for inviting me in 1984 to teach a course on the concept of potential for corporate executives. Teaching this course for several years helped me organize my views on this topic and gave me the opportunity to gather many useful facts from the students about U.S. business.

I am grateful to the graduate students from the departments of Social System Science (S^3) and Decision Sciences at the University of Pennsylvania, who for many years attended my course on the problems in this book, and for their participation in its discussion. In spite of all the difficulties, some graduate students showed strong interest in the course. Graduate students are mostly interested in technical questions because technical knowledge allows them to be competitive in the job market after graduation. Students from the business community, however, are much more interested in a conceptual approach, which would help them think through their work. They are relatively less interested in technical questions either because they have already studied them or because they have encountered them during their work. Besides, these students already have jobs and do not have to convince an employer of their technical competence.

I am grateful to Professors Vadim Arkin, Kenneth Arrow, William Baumol, Igor Blauberg, Roland Dobrushin, Jamshid Gharajedaghi, Xavier de Groote, Larry Hirschhorn, Paul Kleindorfer, Irving Kravis, Joseph Lachman, Efraim Maiminas, Hasan Ozbekhan, Boris Pittel, Viktor Polterovich, Ernest Pressman, Vladimir Rotar, Thomas Saaty, Vadim Sadovsky, Vladimir Shliapentoch, Eric Trist, and Eric Uidin for their useful conversations on the problems of economic indeterminism we had through the years. My gratitude to these scientists naturally does not in any way impose on them any obligations. They certainly could not be accused of expressing unrestrained admiration for my views. The most that I would venture to say is that the majority of my colleagues did not consider my views to be generally known, obvious, or mistaken. These three parameters are important for a negative appraisal of the potential inherent in my concept of indeterminism. However, they are certainly insufficient for a positive assessment of my work. The majority of my colleagues reserved the right to judge the value of my concept at such a time when its results are translated into experimentally verifiable procedures. Unfortunately, while this opinion is probably reasonable within the framework of traditional science, it may not be applicable to the assessment of the incipient stages of a developing concept. This, however, leads to a vicious circle in the assessment of the potential of my concept of potential: it is difficult to accept it as a metaconcept for its own evaluation.

My deep thanks are due to Valery Chalidze for copying the manuscript of the Russian language preliminary version of this book.

Finally, I would like to acknowledge the translators of the present book, Alexander and Gregory Katsenelinboigen, Lena Katok, Constantine Kipnis, Ann Kissin, and Iden Rosenthal and to thank Mary Kane and Marjorie Weiler for typing and correcting parts of the book.

Indeterministic Economics

Introduction

"The time is out of joint." —*Hamlet*, Act 1, Scene 5

BRIEF REVIEW OF THE LITERATURE

In reviewing Kenneth Boulding's *The World as a Total System and Human Betterment*, the well-known American economist Wassily Leontief (1986) wrote:

> It is, however, surprising that the question Mr. Boulding not only does not try to answer, but even fails to pose, is the cardinal question of determinism versus free will — that is, the possibility of real choice. How can the "scholarly community" respond to his call for "normative analysis and the study of human betterment" if Mr. Boulding does not suggest how to solve or, if the question cannot be solved, how to bypass that fundamental philosophical problem?

It seems to me that this is something that could be asked not only of Boulding, but also of economic science as a whole, which deals very little with indeterminism. Because determinism can be viewed as an extreme particular case of indeterminism, a deterministic approach limits the range of economic research; when applied to situations calling for an indeterministic vision, its solutions are far from best. My statement is not intended to negate the accomplishments of contemporary economic science: a deterministic approach is not to be denied, for in many instances its use is justified.

The attempts I know of to explicitly raise the issue of indeterminism in economics are limited and imitative in character, that is, they borrow from other branches of science, particularly from physics. This practice of proceeding by analogy would be justified if the plagiarized field did not suffer from the same poorly developed concept of indeterminism.

Most frequently one encounters the opinion that indeterministic economics has to do with probabilities whereas deterministic economics is associated with complete certainty (Nelson & Winter, 1982). A more intricate version of an indeterministic economy, by analogy with statistical mechanics, is represented as a system where the participants act randomly (Farjounn & Machover, 1983). As a result of their random actions they achieve a stationary process of a certain probability. The authors see the fundamental accomplishment of their work in that they have managed to show that the development of an economic system achieves equilibrium not in terms of certainty of unknown values, but in terms of distribution of probabilities.

This approach to studying economics greatly interests me. Doubtless, some areas of the economy, for example, consumer behavior and the commodities market, could be objects of such a methodology. However, the accessibility of this analogy remains doubtful for the production sphere in those areas where the participants usually do not represent a statistical ensemble and relations between them are to a rather large extent regulated by contracts.

Another approach to indeterminism in economics is connected with the use predominantly of analogies with indeterminism in quantum mechanics (Georgescu-Roegen, 1971; Machlup, 1958).

The latest development of this approach may be found in W. Young (1982). The author actually reduces indeterminism to one particular, albeit an important case that parallels the situation in quantum mechanics. Indeterministic behavior of objects in the last case is characterized by incomplete simultaneous knowledge of both magnitudes of a pair of parameters — the coordinates (primary) as well as the impulse (conjugate).

Young asserts, in the language of economics, that situations also exist in an economy when it is impossible to establish simultaneously both primary and conjugate parameters, that is, the production volume and prices. The author cites several other examples of this kind of situation, such as the simultaneous determination of the level of investment and savings and the level of employment and volume of unemployment. The author justifiably remarks, in generalizing his predecessors in this area, that "In economics, however, the problem involves not only conjugate variables or parameters, but also measurement and adjustment processes which cannot be dealt with simultaneously" (p. 456).

From this viewpoint, he notes, the solution of the problem of indeterminism in economics is connected to the interrelation of the infinite and the discrete, as well as to the choice of the proper lag so that the interval of time, within the framework of which measurement and correction may be made without violating economic functions (for example, the links between demand and prices and its reverse function).

Although the deterministic framework or indeterministic one imitative of physics continues to dominate economic science, the last two decades have witnessed the appearance of wonderful dissident works sketching new pioneering approaches to the problem in question. No explicit mention of the term "indeterminism" appears in these works, but they do present new ideas that have a direct bearing upon the subject matter of the present book. Noteworthy in this respect are the works by Russell Ackoff (for example, 1970, 1981), which explore the multistage approach to development (as opposed to growth and survival), with each stage characterized by its own particular purpose and with a variety of methods used to interlink these stages. This body of work is quite distinct in its methodology from the prevailing spirit of microeconomic research, geared toward attaining optimal growth via a uniform set of methods.

Notably, research dealing with the microlevel — the firm — includes works by William Baumol (1967), Robert Buzzell and Bradley Gale (1987), Richard Cyert and James March (1963), Peter Drucker (1954), Hiroyuki Itami (1987), Janos Kornai (1971), Harvey Leibenstein (1979), Michael Porter (1985), Sidney Schoeffler (1980), Herbert Simon (1982), and others. Common to all these works (especially so in Drucker's pioneering work [1954]), is a strong criticism of orthodox economic theory, which depicts profit as a general indicator of a firm's performance.

Each of the above works seeks to uncover its own particular reasons for rejecting profit as a general indicator of performance, supplementing it with other parameters. Not denying the significance of all these isolated reasons prompting the introduction of indicators other than profit, I want to emphasize a lack of a unifying concept from which all the above causes would stem as particular cases.

Alvin King's (1981) *Index of Vulnerability* is a key to his very constructive criticism of the inadequacy of the profit criterion (although he does not make it explicit) in developing long-range policy on the level of a branch of industry as strategic minerals.

Another group of works expounding an unorthodox approach to economics deals with the macrolevel. Here, we should mention works on *Modern Economic Growth* by Simon Kuznets (1966), on the convertibility of peace to war-time economy by Irving Leveson and James Wheeler (1978), and on reverse convertibility by Seymour Melman (1974). The common thread running through these works is the rejection of gross national product (GNP) as an all-encompassing macroeconomic indicator and the introduction of other parameters reflecting the structure of an economy and its interaction with the environment.

However, all the aforementioned works fail to outline specific methods to commeasure the relative values of these indicators.

Moreover, national income itself is calculated based on the chosen representation of current prices (or prices in a given country). I attempt to show that this approach to measuring national income makes it impossible to measure economic growth unambiguously.

Francis Seton in his work (1985) tried to circumvent these difficulties in using current prices to measure national income and the question he himself posed:

> How can the physical flows and proportions of specified inputs and outputs be compared in rational commensurable terms to produce internally consistent and valid criteria of economic structure and performance, consonant with established notions of cost, use, and value, and amenable to international and inter-systemic comparison without imparting biases inherent in a country-specific, time-specific, or arbitrarily chosen price structure? (p. 4)

In Chapter 9, I shall try to show that Seton was not completely successful in resolving this problem. Perhaps the reason that Seton as well as many other scholars of macroeconomics stumbled over this problem is the lack of a unifying concept that would organically fuse the need supplement national income with other indicators through a modified method of measuring income based on a new type of price.

Work by Richard Nelson and Sidney Winter (1982) occupies a prominent place in macroeconomic research. It places economic disequilibrium as the cornerstone of development and proceeds to investigate indicators, other than GNP, which characterize the state of an economic system from this perspective. This approach lends a more general picture of development of an economic system than the aforementioned macroeconomic research, which advanced reasons for supplementing GNP with other indicators from a more narrow perspective. Another merit of Nelson and Winter's book is an attempt to integrate macro- and micro-operation of the economy. Foregoing a detailed discussion of the proposed approach, I want to note that it represents an important but still a particular solution to the problem in question because it limits the solution to reactive methods. In Chapter 5 I examine the significance and the limitations of the proposed methods.

To sum up, current economic literature seems to lack a unifying (which is not to say exclusive) theory that would

consolidate all the diverse reasons for viewing profit in conjunction with other indicators when evaluating a firm's performance;

assemble all the diverse reasons for supplementing GNP with other indicators when evaluating the performance of an economic system; and

aggregate all the diverse reasons for supplementing income (GNP or profit, for instance) with other indicators when evaluating the performance of the economy as a whole or its individual branches (firm, for instance).

SEVEN QUESTIONS ON ECONOMICS

A theory unifying all these atomized considerations could be constructed within an indeterministic framework of economic development.

What is indeterministic economics, and what distinguishes it from a deterministic one? The questions defy a simple answer, if only because the author himself is a long way from fully understanding the concept of indeterminism. I can only say in my defense that the general category of indeterminism is poorly comprehended. An indirect confirmation of this is the absence of a notion of the degree of indeterminism: determinism is usually opposed to indeterminism with all the intermediate stages ignored.

Lacking a definition of indeterminism, I shall try to immerse the reader into my vision of the problem by posing seven basic questions.

The first question deals with selecting among the alternative courses of economic change based either on technological progress or on the already existing set of tested (and preferably elementary) technologies: in principle, is it possible to judge unequivocally a course of economic development based on technological progress?

The second question searches for order within the chaos of the stock market: is there some kind of order within the chaos of the stock market?

The third question asks if the initial stages of pioneering research in applied science, research aimed directly at some practical problem, yield to outside evaluation. For instance, this question is relevant in judging research on some completely novel method of achieving controlled thermonuclear fusion. Can an outsider (an expert, in particular), rather than the creator himself, judge the advisability of financing new trends in science, even of applied science directed toward a specific and final practical end?

The fourth question deals with the macroeconomic problem of measuring economic growth in current prices: is it possible to derive an index of economic growth based on current prices, in particular by taking prices for a certain period as constant prices?

The fifth question concerns the exhaustibility of profit as an indicator of the firm's performance and the need to supplement profit with other indicators: why is profit, expressed even in probabilistic terms, insufficient in evaluating a firm's performance? Why must profit be supplemented by other criteria characterizing the relations between the firm and its environment?

The sixth question concerns the selection of alternative projects for developing a new technology when we know a priori probability distribution of respective costs and benefits: when faced with several alternative projects (described in probabilistic terms) for developing a new product, is parallel implementation ever warranted?

The seventh and last question spotlights the connection between the various ways of representing an economic system. One method of description — analytical or by means of a tabular function — presupposes reactive behavior; another method — equilibrium or optimal models — entails selection. For instance, can we say that an analytical representation of the consumer demand function is linked with a deterministic approach whereas an optimal model is linked with an indeterministic one?

Addressing these questions constitutes the core of the book. The fourth and fifth questions are examined in greatest detail because they call for new, less studied methods of economic analysis. The other five questions are intended to present a more complete picture of economic development by illuminating the issues raised by questions four and five.

I now want to make some brief general comments regarding the two questions on which I focus.

At times in the history of science, a powerful new theory is created that explains the overwhelming majority of phenomena within the realm of its purview. At the same time, some phenomena seem, to many professionals in the field, to lie on the periphery and thus seem not as important. However, the prevailing theory is powerless to explain them, despite energetic attempts to squeeze these phenomena into its framework. This could mean that the existing theory is only an important special case of some more general theory. On the basis of new axioms, the more general theory not only could explain those phenomena explainable by the old one, but also could go further by solving the aforementioned peripheral problems as well.

One famous example of this situation comes from mathematics; it is the problem of Euclid's fifth postulate (Smilga, 1968). For more than 2,000 years, mathematicians could not answer the seemingly secondary question: was the assertion in Euclidean geometry that there exists a unique line parallel to a given line and passing through a point outside that line a postulate or a theorem. In the nineteenth century several mathematicians came quite close to resolving this issue, realizing its full implications. Among them was the great Carl Friedrich Gauss (1777–1855). However, fearing ridicule, he decided against publishing his results on this topic; they became known much later from his notes and letter to colleagues. A contemporary of his, who also contributed to solving the problem of the fifth postulate, was the Hungarian Yanos Boyai. He published the results of his investigations as an appendix to a mathematics textbook written by his father, Farkos Boyai. An outstanding Russian mathematician Nikolai Lobachevsky (1792–1856) made a great contribution to solving this problem. His work on the subject was published in 1826. His contemporaries not only failed to comprehend it, but even ostracized it. The decisive blow to the problem of the fifth postulate was dealt by the discoveries of an eminent German mathematician Bernhard Riemann (1826–1866) in the field of non-Euclidean geometry.

The solution of the problem of the fifth postulate, which seemed a minor scholastic issue, had a revolutionary impact on mathematics. Eventually, these advances in mathematics exercised great influence upon physics, shaping in particular the mathematical apparatus for Albert Einstein's relativity theory (1961).

The theory of relativity and its relation to classical Newtonian mechanics could serve to illustrate the effect of minor peripheral problems on the development of science in general. It was precisely these seemingly peripheral issues stemming from having to explain the puzzling orbit of Mercury on the basis of classical mechanics that were largely instrumental in the development of the theory of relativity.

I see certain parallels between the current state of economics and the just described situation in physics and mathematics. The analogy reflects the fact that economic science has accumulated and even systematized a large amount of empirical data. However, this experience has not been properly conceptualized within the framework of modern economic research. The problem of measuring economic growth serves as a partial confirmation of the point I have made.

Over the last century, economic theory has made great strides by embodying many of its representations in mathematical models. I would like to stress here the importance of linking prime and conjugate economic parameters corresponding to real and monetary terms respectively. These models, although to a large extent analogous to those in physics and having physical analogues (Rozonoer, 1973), have at the same time paved the way for significant advancement in interpreting value parameters. Mathematical programming has managed such progress because of economics. Development of mathematical tools, linear programming in particular, for the analysis of economic problems carried out by George Dantzig, Leonid Kantorovich, Tjalling Koopmans, and John von Neumann has been accompanied by economic interpretation of the various value parameters employed in the solution procedure. Research in this area has enabled us to find noncontradictory methods to form values as instruments for the autonomous operation of economic units, each possessing degrees of freedom in decision making while unable to overview its surrounding environment directly. In this area economics seems ahead of the other sciences. Its models are quite instructive both for physicists, sociologists, biologists, physiologists, and other specialists dealing with the organization of self-acting autonomous systems.

Meanwhile, on the periphery of economics a question remains open: what system of values measures changes taking place in the economy over time (by the same token, this problem has a spatial aspect in comparing the level of development of different countries)? This is precisely that nagging fourth question posed earlier: how is the growth index for the economy to be computed if the output of various goods and their respective prices vary disproportionately over time?

Let me recapitulate the essence of the question for the uninitiated reader. Assume the country produces two goods — A and B. A denotes capital goods (which might include trucks, generators, and steel) and B — consumer goods (such as bread, houses, and televisions). Take two moments of time, say year 1900 and 1990 (denoted as T_1 and T_2, respectively); being rather remote from each other, this period of 90 years might have witnessed significant changes in the output and prices of these two commodities. The problem is to calculate the overall growth of production over the period in question.

Economists realize that unlike commodities (eggs and steel) cannot be added directly. Thus, it becomes necessary to find a criterion that would reduce all these products to one common denominator. Prices seem to qualify as such a criterion. But prices of different goods vary differently over time, so base prices were proposed in order to eliminate the impact of price changes. This solution breeds new problems.

As Table I.1 indicates, difficulties arise because the growth index depends upon our choice of the base year; that is, the index is not invariant with respect to base prices. In our example, the growth

TABLE I.1
Production Growth Index

	Years				Total Production			
	T_1		T_2		In Prices T_1 of Years		In Prices T_2 of Years	
Products	Output	Price	Output	Price	T_1	T_2	T_1	T_2
A	5	30	80	60	150	2,400	300	4,800
B	20	6	22	24	120	132	480	528
Total					270	2,532	780	5,328
Index					100	938	100	683

index equals 938 if prices at T_1 are chosen as the base but only 638 if they are chosen at T_2.

Economists persist in trying to circumvent these inconsistencies. Even though they have made significant progress in a number of specific cases, the general solution is still evasive, and, as far as I know, no new paradigms are in sight.

"Well, all right," a reader will say. Suppose economic science is helpless in answering the question of measuring the growth index. Still, it is an isolated question borrowed from macroeconomics and seemingly completely out of tune with the ones that precede or follow it, such as the fifth question, which deals with the role of profit in guiding the firm's development.

Just to remind the reader, the fifth question "doubts" the sufficiency of profit as the criterion of evaluation of firm performance. The leading theory expounded in the economic literature states that the firm's ultimate objective is profit maximization. This approach is not limited to theoretical research or mathematical economics in particular; it permeates most microeconomic textbooks intended for students at a level as advanced as business schools. Of course, because the interaction between the firm and its environment are subject to a variety of random fluctuations, many works qualify the statement concluding that profit maximization should be sought in terms of probabilities.

But apparently, businessmen's decision making often incorporates, along with profit, other indicators that reflect the relations between the firm and its environment (or those arising within the firm). An example of one such indicator is the firm's relative market share. A very natural question arises: "Why should relative market share be considered along and, in fact, on par with profit?" It seems that profit encompasses all aspects of the firm's performance and its interaction with environment. Optimal market share is then a function of the maximum level of profit. In fact, putting a firm's relative market share on par with profit gives rise to some nasty problems of commeasuring these two sometimes conflicting indicators because the growth of relative market share may require sacrifice in profit or vice versa.

A reader even slightly familiar with the workings of an economic mechanism will immediately respond with general and, for the most part, sensible arguments in favor of evaluating relative market share independently and on par with profit. But he would be stumped if asked to explain the proper amount of a sacrifice in profit in favor of increasing a firm's relative market share.

Thus the prevailing approach to firm performance, expounded in the literature on economic theory and mathematical economics, in particular, operates under the assumption that the criterion of firm performance is profit. For the most part, this category of economic

literature eschews a conceptual explanation of the very critical question of why it is necessary to introduce criteria other than profit (for instance, market share) in evaluating a firm's performance. I want to point out in this connection that there is a gap between economic literature and business literature, which does touch on this topic.[1] However, the latter also lacks a conceptual framework within which to study the problem, thus distancing itself from the real-life work of successful business leaders, who must solve these perplexing problems intuitively. In fact, on those occasions when this far-from-perfect literature on business does make its way into business practices, individuals who are not sufficiently critical of it can find themselves in trouble.

To sum up, I propose that the answers to the fourth and the fifth, as well as the other five questions posed above, hinge on one and the same poorly elaborated two-fold problem: introduction of the notion of a degree of indeterminism and development of diverse methods of economic operation, corresponding to different degrees of indeterminism in the system. The answer to questions four and five hinge on the introduction of a new phase in the indeterministic spectrum and adequate methods of operation corresponding to this phase. Discussion of this new phase and its respective methods comprises the core of the book.

SYSTEMS VISION OF THE BOOK

In keeping with the systems approach, any problem involves five quite independent aspects, namely function, structure, operation, operator, and genesis, all integrated.

Functionally speaking, the present work is an attempt to answer the seven questions posed above, focusing on questions four and five and showing that essentially we are dealing with the same problem: multiphase representation of degree of indeterminism.

The above considerations regarding the functional role of the book must be bridged with the book's mode of consumption. A book has at least two modes of consumption. In one case, the work aims (its mode of consumption) at providing techniques that, if imitated by any reader, would lead (under the same initial conditions) to the same result; that is, it purports to objectify the proposed methods. Ultimately, this approach might lead to procedures being formalized to such an extent as to be handled by computers. This approach predominates in teaching students at business schools.

Alternatively, the work strives to enrich the intuition of the reader. The reader is presented with a concept that should help him realize what he might be doing anyway, but intuitively and so perhaps not in the best possible way. A framework is constructed within which the reader is expected to improve his own methods of thinking by discovering some hereto unknown techniques. This approach does not negate intuition but enriches it leaving plenty of room for intuitive activity.

The two approaches should not be opposed to one another. The greatest benefits are derived when they are combined. With the leading role delegated to the enrichment of the intuition, the use of objective procedures for routine tasks (perhaps delegated to computers) might be highly conducive to creative activity. The engineering field today is an example of this situation.

The structure of the book consists of 17 chapters, grouped into 3 parts. The opening part of the book mainly discusses the methodology of defining degrees of indeterminism in an economic system. The groundwork for my vision of this problem is the three phases of indeterminism and the methods of operation of the economic system corresponding to these phases. These methods I call directing, potentiating, and programming.

The major part of the book, Part III, concerns the method of potentiating. The core of this part is made up of one chapter, Chapter 8, dealing with the essence of potentiating. All other chapters of this part are corresponding to the application of the potentiating method to the macroeconomic and microeconomic levels and to an intermediate level — a branch of an economy.

Preceding this major part is Part II, devoted to the other two methods of systems operation within an indeterministic framework, namely directing and programming methods. Violation of the established sequence of methods — skipping from directing to programming — was intentional on my part; I decided it would be more expedient to introduce the reader to my approach step-by-step by first examining the polar cases of the indeterministic spectrum.

The material used in this book pertains first and foremost to Western economies. However, I have made use of material pertaining to the former Soviet Union. Although the USSR is no more, much of what has been said holds true at least for the leading former republic, namely Russia.

My methodology — the operational (process-oriented) aspect of the analysis of the nature of an economic system — is founded upon the philosophical ideas of indeterminism. I believe that systems approach can shed new light on this old category.

As noted by Russell Ackoff (1981), systems approach is characterized by expansionism (rather than direct analysis of a system); that is, immersion of a given system in a more general system — a suprasystem.

A system can be analyzed from a multilateral perspective — an approach elaborated by Jamshid Gharajedaghi (1985). The many facets of a system include: the name of what — function; relation between the final product and the elements required to make it (including energy and information) — structure; process of transformation of primary ingredients into final product — process; tools used to implement the process — operator; and the impact of the system's prehistory on all the aspects mentioned above — genesis. In principle, all these aspects are independent. This means that if one aspect is fixed others still maintain a certain degree of freedom. Systems vision consists of multifaceted analysis. Unnecessary arguments among scholars in the same field frequently arise as a result of some system being viewed from a single narrow perspective.

If I had to specify the general approach of systems methodology, I would venture to say that the basic principle is that of design. Discovery is really a result of collating the set of constructed models with one model, which, based on all the accumulated experimental data, is taken to be the model of the external (relative to the constructed models) world, the so-called real world. This approach is reminiscent of Oscar Wilde's aphorism that art is not a reflection of life, but life is a reflection of art. This does not repudiate the role of the so-called real world in the creative being, both in stimulating ideas for creation and in testing their validity in the so-called real world (an experiment), as well as in their actual realization.

The operator implements the above mentioned methodology. In our case, the operator is the author of the book, armed with a modern computer. What does the operator have to show to justify his ambitious project, expressed in the title of the book? Books I have written on the theory of optimal planning, my more applied research dealing with effectiveness of technological progress, and my contacts with the real-life economic practices, both in the Soviet Union and in the United States, give me reason to think I am professionally qualified to deal with economic issues. My long-standing interest in philosophy, corroborated by my attempts to develop a new axiological concept, reveal my inclinations to look at the age-old problems of philosophy in the new light.

It seems my personality fits well with my ambitious project. I recall the summer of 1953 when I took part in a study to determine the reserve capacities at a ceramics factory. After two weeks at the site, I discovered a number of inconsistencies in the instructions and was quick to report them to the leading engineer, a rather old and fearful Jew, Moses Taub. His response was brief: "You are an extremely immodest and ambitious young man. Do you really think that all the authors and all the reviewers of the instructions have overlooked these deficiencies? Do you really think they are more stupid than you?" My answer was affirmative. I added that "modesty spells doom for the young scholar." Perhaps my response was lacking in tact. A former inmate of Stalin's "penitentiary" system, Solomon Heinman, enjoyed saying youth should "dare but not defy!"

Whatever I say, it is up to the reader to judge how adequately my ammunition has supported my ambitions.

Finally, I want to touch upon the genesis aspect of the book, that is, its history.

I hope the reader will take time to review the essential aspects of the problem. For my part, I will try to set out my initial perception of the problem that, about 20 years ago, impelled me to write this book.

I shall begin at the somewhat intermediate stage when I first learned of the problem (actually, I became acquainted with it during my student days from a course on industrial statistics taught by Semen Iugenburg) but did not understand its specifics. Moreover, at this stage, which dates back to the 1960s, I was corrupted by the then pioneering economic theory of optimal planning. Optimal planning assumes that the criterion of development of an economy is known and can be linked with the existing constraints in the form of technologies and available resources. I presumed in this case all the answers to the questions concerning the dynamics of an economic system could be derived in the course of solving a corresponding optimization problem. Optimality criterion reaching its highest value under given constraints means that the optimal trajectory has been found. The problem of measuring the rate of growth of production in real terms in a system of optimal planning simply does not make sense: change in the value of the optimality criterion characterizes economy's development. I assumed the same applies when comparing the levels of development of different countries. My starting assumption (whose dubiousness became clearer to me much later because of my discussions with Victor Polterovich) was that all people in all countries want the same thing: they seek to increase their economic prosperity. In this case the value of the optimality criterion for different countries should reflect their relative level of development.

Meanwhile, many economists, including those of the old conservative school, believed that measuring the rate of growth of one country's economy or comparing the levels of development of two different countries represents an independent problem. I had conflicting opinions on this point. I first thought that the economists who raised this issue were old-fashioned, they did not understand that the problem simply vanishes with the introduction of the new concept — the optimal planning theory. At the same time I was also willing to concede that optimal planning theory could be missing something in its assumption and that, if so, the independence of the problem of calculating the rate of growth would be affirmed. Of course, even then the problem would have to be solved within the framework of a certain concept. But this realization came to me later.

A landmark in the evolution of this new concept occurred when I realized the connection between nonoptimal planning and the problem of measuring economic growth. The initial impulse was the perceived analogy with positional style in chess. It required a new calculus based on the unconditional values of material and positional parameters as independent variables.[2]

Generalizing the issues involved in chess (and social issues, military science, theology, and other realms of human endeavor, as well as biological systems), I reached a conclusion that the answer to the question of measuring economic growth reaches to a more general problem of measuring economic development. But this calls for a reexamination of the fundamentals of economic science, making explicit the assumptions underlying its indeterministic nature. Here, new problems of understanding the economic mechanisms operating under the conditions of indeterminism arise. Issues raised here are outside the scope of orthodox trends in economic science, which are based primarily upon deterministic notions and respective models of equilibrium (optimality).

So, at the beginning of the 1970s, the concept of economic indeterminism began to take shape in my mind. I have continued my research after emigrating to the United States. Lectures to graduate students and professionals on the subject and conversations with my colleagues have helped me elaborate the new computational techniques and define their proper place in the pantheon of other works dealing with the same subject matter. My deliberations materialized in a number of conference papers and in two books (1984, 1989) devoted to a variety of problems concerning indeterminism.

NOTES

1. The list of authors seeking to inject new ideas into microeconomic theory includes both economists and scholars of business (including management and organization theory). More often than not, economists tend to ignore business scholars' research and vice versa. Suffice it to say that *Business Behavior, Value and Growth* (both the 1959 edition as well as the 1967 edition) written by outstanding economist William Baumol fails to cite or even mention the notable scholar of business Peter Drucker, who had in 1954 published the classic *The Practice of Management*. Although the two books differ considerably in terms of formal presentation of material, they share one fundamental idea — strong criticism of profit as a sole criterion of a firm's performance. Not to be outdone, a book by a well-known business scholar Michael Porter (1985) lacks a single citation of William Baumol (1967). This in spite of very close symmetry of the two books in opposing profit (each from its own perspective) as a sole criterion of firm performance.

2. I want to underscore the extensive parallels between decision making in chess and in an economy. Herbert Simon (1982), a Nobel Prize Laureate in economics and a chess expert, has extensively explored the cross-fertilization of the two fields. Rudyard Istvan (1984), the vice president in the Chicago office of The Boston Consulting Group, brought analogies between chess programs and economic programs. The former chess champion of the world Michail Botvinnik published an article in the newspaper *Pravda*, March 30, 1981, concerning usage of his original chess computer program for scheduling power plant repairs.

I

ECONOMIC SYSTEM AND INDETERMINISM

1

Economy as a Dynamic System

DEFINITION OF AN ECONOMIC SYSTEM

If one would like to avoid the difficulties of this way of defining the system, then he may respond to the question, "What is the given system?" by paraphrasing the mathematicians' answer to the question, "What is mathematics?" "Mathematics," they respond, "is what mathematicians study." Thus economy is that which is studied by economists.

Another definition of an economic system that follows a thornier path arises from the methodology of the systems approach described in the introduction to the book.

First of all, there is no "given once and for all" definition of an economic system. Attempts to formulate one ultimate definition give way to a pluralistic mechanism of search for the most relevant definition (Katsenelinboigen, 1989). The gist of this search mechanism is the following: the initial stage involves the construction of a possibly expanding set of definitions. Subsequently, specific investigative objectives determine the choice of the most pertinent definition from the set. If the definition is found lacking over the course of a given investigation or the objectives of the analysis shift, then the employed definition is replaced by another one borrowed from the above mentioned expanding set of definitions.

Then, following the systems principle of expansionism, one could generalize the definition to delimit a suprasystem that embraces a given system.

The suprasystem can be arranged in two ways, functionally or structurally.

Immersion of a system into a functionally organized suprasystem explicitly distinguishes a set of implied elements comprising the system. Our particular case spotlights the fact that an economic system is a system whose participants are people.

In principle, economic considerations apply not only to mankind, but also to animals, insects, and even to the realm of self-acting systems in the inorganic world (Heinrich, 1979; Pryor, 1981; Rozonoer, 1973). We shall consider a subset of definitions that enlist people as participants of an economic system.

In our particular case, immersion of an economic system into a functionally structured suprasystem entails the discernment of the various interrelations between the economic system and the social system and nature that interact with it directly. Our subsequent analysis is predicated upon this implied interface between the economic system and its suprasystem.

The next step is to employ the systems approach to construct a set of plausible definitions of an economic system (limited by the above qualifications). The constructed definitions ought to take the following aspects into account: function, structure, process, operator, and genesis. Every definition of a given system can be formulated from a different perspective.

One can concoct a rough definition of an economic system and then compare it with the existing definitions. This approach might reveal certain deficiencies of the existing definitions as well as rectify those of the newly devised ones.

The method that I picked is different. I began with a critical examination of the prevailing definitions of economics from the standpoint of the above mentioned aspects of a system. In the end, I arrived at one possible generalized definition of an economic system. That most of the existing definitions incorporated just one of the relevant aspects justifies this approach.

Now, the systems methodology directs us to pick one (or several) definition from the diverse set of previously constructed ones. In case we are dealing with several definitions, each definition must be inspected in terms of some particular aspect of analysis in order to distinguish certain features peculiar to it. We can then construct a new definition that represents a combination of the discovered characteristics.

However, I was deprived of this choice, for I am not familiar with any attempts described in the literature to construct a definition of a given system following the charted method. As a result, I can only advance my own definition of an economic system. So, let us embark upon an analysis of the existing definitions from the various perspectives of the systems approach.

From the functional point of view, *Encyclopedia Americana* (1986) attempts to define economy as a system that produces and distributes material goods, energy, and services. "Economic systems are forms of social organization for producing goods and services and determining how they will be distributed." But what about such services as national defense? Is the direct output of this area military or economic? Perhaps the economy is restricted only to the production of goods directly necessary for survival in our struggle with nature? How then to delineate these goods? To what degree is music necessary for survival? Is it not sometimes more valued than material goods?

From the structural point of view, the economic system is defined by the *Great Soviet Encyclopedia* (1978) as the totality of all the sectors of production. "Economy. National economy or some part of it incorporating particular branch(es) and modes of production."

From the process-oriented point of view, the economy is understood as a system in which limited resources are allocated (Moffat, 1976). In many economic publications this definition of the economy is not given directly, but rather via a definition of the theory that studies it, that is, through a definition of economics. "ECONOMICS. Most definitions are similar to: examining the allocation of limited resources for the satisfaction of UNLIMITED WANTS of man" (p. 94). But people are concerned with the distribution of limited resources in all their activities. Therefore this can scarcely be considered the hallmark of economics.

Kenneth Boulding (1970) sought to give a more specific operational definition of the economy, stressing the method of resource distribution. This he considers to be exchange: "the economy of the econosphere . . . consists primarily of that segment of the sociosphere which is organized through exchange and especially commodity exchange" (p. 11). The exchange of goods, as a particular case of exchange, according to him assumes relations under which "I will give you something that you want if you will give me something that I want" (p. 10).

In this instance only the market represents the economy. At the same time Boulding (1970) notes that each organization in society, even if its internal operation is based on nonexchange principles, is an economic actor inasmuch as it participates in the exchange of goods. In that sense a centrally planned system is an economic system only to the extent that the state represents one of the participants in relations of exchange within as well as outside the country.

But condition of exchange is not sufficient to define an economy. Relations of exchange have their place in many other aspects of human activity as well. Thus in a notable number of sociological works, social relations are interpreted as relations of exchange (Hingers & Willer, 1979). The detailed particularities of relations of exchange must still be elaborated in order for them to satisfy the demands of definition as an economic system.

According to another work by Boulding (1968), relations of exchange can be considered economic if they satisfy the necessary condition of being "relatively clear and measurable." But even with this more limited concept of exchange the definition of an economy proffered by Boulding seems to me to narrow the class of economic systems because it is limited only to market-type mechanisms.

One could generalize the concept of economic relations stemming from interactions among specialized participants and these relations as "relatively clear and measurable" transfers of goods independent of the social relations resulting therefrom. In this case all the same processes that in Boulding are accomplished via horizontal mechanisms (markets) could just as readily be handled by vertical mechanisms, most particularly by centralized planning making heavy use of such institutions as prices and money.

An economic system is sometimes (*Encyclopedia Britannica*, 1983) defined from the same process-oriented point of view as the aggregate of all institutions that organize economic life. "An economic system may be defined as the sum total of institutions and patterns of behavior that organize economic activity in society." Taking this into account, it is possible to introduce the concept of an economic mechanism and say that it is characterized by the sum total of institutions that are based on the type of economic relations defined earlier.

As far as the operator aspect of defining an economy is concerned, that is connected with the emphasis on the substances used to carry out economic processes, that is, natural resources, capital, and labor. Thus, in one definition of an economy (Farjoun & Machover, 1983), the stress is placed on human labor as the decisive particularity of economic processes.

> Labour is, *par excellence*, the essential substance of an economy, and should therefore be taken in economic theory as the fundamental yardstick. . . . The science of economics, taken in the most general sense, is concerned with the study of the social processes and structures by means of which and through which social labour is organized and performed, and the output of this labour distributed and allocated to various uses (p. 85).

I do not wish to negate the enormous role that labor plays as a factor in economic processes. However, its exaggeration and the underestimation of the role of the other factors can lead to the construction of economic theories that prove internally contradictory. This observation is completely applicable to the labor theory of value. It is no accident that the authors of the above definition of an economy seek to reinstate this theory using rather modern analogies from physics.

Finally, let us turn to the genetic aspect of defining an economic system, although by itself it apparently has no impact on this definition. At the same time it shows that a definition of an economy does not have to be associated with a rigid, hardened system. As economic relations in the sense indicated above broaden, so the area encompassed by our definition of an economic system expands. For instance, the realms of education and medicine were long regarded as outside the scope of economics. The creation of such disciplines as "the investment in the human capital" assigned this realm to the potential sphere of economics. The orbit of economics is constantly expanding to include the multistage sphere of R&D. At one point, when science and invention were the task of a few scattered individuals, when the time lag between basic science and its practical implementation was considerably longer than today, economic science concerned with the economy proper tended to focus on the application of the inventions in the production of final products. Today the R&D sphere is fast becoming subject to economic analysis. This shift is evidenced in the work of Edwin Mansfield (1977), a well-known expert in the economics of innovation.

In summation, an economic system can be defined as that part of the human social system in which the production of goods (material as well as nonmaterial) takes place within the framework of limited natural resources, labor, capital, and knowledge, with the economic mechanism predominating. In this connection it is important again to underscore that the economy as a system cannot be reduced to an economic mechanism. It includes a plethora of various mechanisms, such as administration, but the economic mechanism predominates.

NORMAL AND PATHOLOGICAL ECONOMIC SYSTEMS

Development of an economic system requires an appropriate mechanism of performance; without it the system is doomed to disintegrate. This brings us to a rather interesting classification of economic systems — healthy and sick on the one hand, and normal and pathological on the other.[1]

A healthy system is a system that operates according to the will of the people, provided, of course, they are not oppressed by the authorities. A sick system is a system that deviates from the will of the people. No doubt this condition is very difficult to verify, but let us for the moment assume that the distinction is valid. In some cases this requirement can be verified; for instance, if the system decreases the amount of material goods while people seek an improvement in their material well-being, this system could be considered sick.

Another dimension is the normality-pathology duality. Pathology, as manifested by socioeconomic systems, can be defined (Gharajedaghi, 1985) in the following way: "The inability or lack of desire in a government/management of a social system to remove a persistent obstruction to development that can be removed by a change in either (a) the way the system is organized and/or (b) its social or physical environment" (p. 71). What distinguishes a pathological system from a normal one is the inability of the former (within the means available to it) to maintain normal operation and handle adverse situations. Sickness is not equivalent to pathology, which refers to the inability of the system to treat its ills.[2]

The two dichotomies mentioned above (healthy-sick, normal-pathological) give rise to four possible combinations, only one of which is logically inconsistent — an economic system that is at the same time both healthy and pathological.

Let us extend this classification of economic systems to capitalist and socialist systems in particular. In terms of the classification given above, the Marxist view of capitalism is that it is both unhealthy and pathological. Capitalism is plagued by such ailments as economic crises, unemployment, homeless people, inflation, prostitution, and crime that cannot be rectified within the system and condemn it to failure. Marxism generally regards the socialist system as being both healthy and normal.

Of course, this view of social systems is rather radical. Let us add a few shades to the Marxist black and white picture of the world and try to pinpoint both advantages and disadvantages inherent in these two types of systems. Under this more moderate approach (still within a Marxist framework) the flaws of capitalism — its ills and pathologies — would eclipse what could be called its healthy and normal features. Under socialism, the reverse is true.

This view of the world reached its peak of popularity in the 1930s. The deep economic crisis that engulfed developed capitalist countries in conjunction with World War I, which preceded it, and a general degradation of morals were a result of the development of capitalist countries and can in no way be attributed to any communist plot. At that same time the Soviet Union was starting a program of rapid industrialization with mushrooming cities, a deficit of labor power, and optimistic songs sung by millions. That all these "accomplishments" were accompanied by the destruction of a strong peasantry and a subsequent bacchanalia of political trials over the so-called "enemies of the people" was viewed by liberals as some unavoidable by-products associated with growth. In other words, liberals saw capitalist ills as a confirmation of the pathological nature of Western societies whereas those suffered by the USSR were considered temporary evils associated with growth — ills of a normal society. In retrospect it is easy to see how difficult it was for liberal intelligentsia to understand which structure is pathological and which is normal. Also all the afflictions plaguing the Western world were exposed in the free Western press, which only aggravated the apocalyptic mood aroused by current events. At the same time, the Soviet press was subject to rigid control by the party apparatus. Deaths of millions of people were tightly concealed, so extreme liberal circles may have even considered news of genocide in the Soviet Union as a hoax fabricated by the enemies of the new regime.

Now in retrospect 75 years of history have shown which system, from the standpoint of a Western liberal, is normal and which is pathological. The Western world has managed to overcome at least the critical stage of its diseases although they are still present and continue to cause much distress. Soviet society has fallen into an era of stagnation.

BREAKS IN THE ECONOMIC NETWORK

Creative Nature of Processes Unfolding in the World

Economic literature is replete with notions of the economic system as, primarily, a system based on a fixed set of products and the means of their production, which can function within the framework of complete and noncontradictory mechanisms. Meanwhile, the economy is a dynamic system. It is host to processes of evolution related to creation of new products and the means of their production; at the same time the processes of evolution do not in themselves have to be complete and noncontradictory. Study of the processes of evolution in the economic system can be aided by their perception in the general context of the world's development.

Metaphysically, one may approach the dissection of the world around us through the premise of its perfection in the sense that there is completeness and noncontradictoriness in the mechanism of its functioning. In this case, imperfections in the world are merely the result of the imperfection of our epistemology of comprehending the world; hence, the researcher's task consists of finding the kind of epistemology that would allow the discovery of the world's perfection. That was precisely what characterized Leibnitz and Einstein where the physical world was concerned. An analogous approach has often been suggested during the construction of new socioeconomic systems. Notions of human perfection and the existence of paradise on Earth, coupled with the extremely imperfect organization of existing human society, nurtured the creators of socioeconomic utopias, who claimed to use the scientific method in dealing with the development of society.

Metaphysical notions of the perfection of the world around us can turn out to be heuristically precious for the formation of quite important conceptions. Einstein's theory of relativity is the best example of this. At the same time, considerable dangers lurk within these given notions, particularly when they are applied to the world around us as a whole. They often create the illusion that the world's perfection can be discovered with one last stroke and then all problems will be eliminated forever and ever. The oft-proclaimed completeness of physics or the dream of building a communist society can serve as examples of attempts to discover such a completed notion of the world.

Examination of the world can also be approached on the basis of the metaphysical premise that the world around us is imperfect but evolving. Here, in turn, two extreme viewpoints are possible. According to the first, the world's mutability can be attributed to the fact that within the framework of extant methods of interaction not only are the known structures repeated, but also, due to accidental breakdowns in these methods and external accidental changes, accidental new objects are created. At the same time, scant attention is paid to the changes of methods of interaction on the basis of metamethods of changing these methods.

According to the second extreme viewpoint, a creator completely regulates the mutability, that is, completely and noncontradictorily accords what is being created (new and old) with what had previously been created. It might even be recommended that this regulating not be exposed, for the ways of the creator may turn out to be inscrutable.

However, one may admit the premise that between these two extreme ways of changing the world there are internal evolution mechanisms of various types, leading to the creation of the new, but with incomplete and contradictory regulation. In other words, in the world there are always ongoing changes and creation of the new, and this mutability will always contain discrepancies — incompleteness and contradictoriness. These discrepancies are only removed in certain relatively enclosed portions. The ever increasing participation of physicists, and inorganic scientists in general, in the study of the evolution of the inorganic world promises significant discoveries in this respect. A global theory like the Big Bang, as well as specific works dealing with various portions of the inorganic world, for example, the evolution of minerals (Zhabin, 1980), constitutes attempts to probe for new approaches to the development of the inorganic world. Of course, methodologically speaking, these attempts are based on deterministic notions of complete and noncontradictory cohesion of the world. Were we to speculate, it would be admissible to discuss the mechanisms of creativity in dealing with the inorganic world akin to the creative evolution of the living world, according to Henri Bergson (1911), on the whole and of human society, in particular.[3]

However, the relative slowness of the unfolding of processes in the inorganic, and even in the biological, world hinders their

comprehension as creative processes. In the economic system the evolution proceeds incomparably faster and more visibly, and it is, therefore, much easier to discover all sorts of complexities of creative processes there. Everything said above leads me to the idea that the dynamic world of the economic system can be viewed as accelerated photography of general evolutionary processes in the world around us. This approach feeds the hopes that economics will not just be, for the most part, a consumer of ideas from other sciences but that, in turn, ideas derived from economics will to an ever greater extent fecundate other sciences (Tugan-Baranovskii, 1973).

Needless to say, such metaphysical notions of the basis of evolution of the world are an act of faith. They are quite speculative as far as the physical world is concerned; they are considerably more substantiated for the biological world, and will probably prove to be justified, for the most part, for the economic world.

Below I would like to explore some of the causes of the appearances of gaps in the ties of the economic system with its suprasystem and, to an even greater extent, within the economic system, moving from the premise that economic systems are characterized by omnilateral mutability both with respect to external resources and the forces guiding it and to the internal changes taking place within it.

External Breaks of the Economic System

First of all I would like to touch on the changes in the guiding forces of the economic system, that is, the predispositions of people in time (as well as in various countries).

These guiding forces will be distinguished on the basis of whether in a given society the resolution of human problems is sought through the change of the environment — extravert culture — or through changing the individual's inner world — introvert culture. However, no matter how the guiding forces defining economic dynamics may be assigned, there remains the question whether they can be tied together into a single complete and noncontradictory network. This is relevant for both the past and the future. In dealing with the past, it remains unclear whether such networks can be constructed with the information available. That there is no single complete and noncontradictory conception of the economic history of humanity is testimony to the difficulty of constructing such a network. But even if such a network could be successfully constructed, when considering the conditions of equifinality it would be far from saying that this network necessarily reproduces what had taken place. At issue is only the creation of a complete and noncontradictory system. Furthermore, can a network like that be constructed for the future? Is it possible to give a forecast for the entire boundless future and to possess the decisive rule that would guarantee that changes will proceed only in a given direction? Is it possible to at least know all events ahead of time far enough (or to possess the decisive rule guaranteeing the direction of changes) that, no matter what events take place beyond the visible horizon, they will not exert any influence on our behavior in the period limited by this horizon?

Thus, in the conditions of people's predispositions changing in an uncertain manner and their changes of values of various benefits, be they spiritual or material, aimed at military might or improving the standard of living, it is apparently difficult to construct a single network of changes in human values eternally applicable.

Everything said about the absence of complete regulation in the changes in human values is also relevant to changes stemming from another external source — nature. At the same time the latter changes can be even more unregulated and unexpected because it is not clear how the eternally changing world can influence the inputs of the economic system on our planet.

However, I would like to focus on the internal changes in the economic system related to its scientific and technological aspects (including organizational changes among the latter). This intention is motivated not just by the enormous role of these processes in the development of the economy; it also possesses important general methodological significance.

INTERNAL BREAKS IN THE ECONOMIC SYSTEM

The Two Sources of Development of the Economic Network

Our analysis of the role played by the economy's inner processes in instrumenting complete and consistent reticulation of its subdivisions follows an outline and incorporates a number of highly simplified assumptions.

Initially, an economic system could be represented as a system in which the production of final goods — from consumer goods to military means — is carried out directly. This means that a person, using his own body, wrests from nature the products he or she needs. Remnants of this kind of process still survive — I refer to the collection of wild berries, mushrooms, and the other gifts of nature. This kind of system of production consists of a number of parallel links, brought together in the person of the individual worker. Thereby a link (Ashby, 1956) is a unit of the economic system comprising the what, the from what, the by what methods, and the operators (the combination of man and tools) — all part of the process of transformation.

As man developed, or perhaps at the time of our very origin, multilink production chains formed between nature and the final product. They are characterized by increasingly refined technologies composed of many stages in the processing of nature's material and the creation of many intermediate products before the final good is obtained. These chains were engendered primarily by the production of tools, among these second-order tools, that is, tools for making other tools.

However, the production chains thus formed were in many ways isolated, parallel, and comparatively short. Making both the tools used in them and the intermediate products versatile transformed this mass of production chains into a unified network.

Inasmuch as this network was at the start sufficiently transparent, a change in any one of its links could be directly connected with the results in terms of final production. As such networks grew, became more complicated, and there appeared in each link degrees of freedom as to the production of different goods and consumption of different resources, it became no longer possible to directly relate the changes in each of the links with the results. In response to this, more refined methods were created enabling us to measure indirectly the influence of changes in any arbitrary link exerted upon the final outcome. This kind of projection allows for an autonomous development of a given link. In other words, it enables a given link to select from its feasible space of alternatives the path that comprises a constituent fragment in the chain of effective development of the network as a whole. Subsequent integration of this complex economic network is conducted via horizontal and/or vertical mechanisms. At this point the actual mechanism (horizontal, for example, market, or vertical, for example, planning) that integrated the network is completely irrelevant. The existence of invariants is crucial. One such invariant is the institution of prices. Prices guide the movement of each separate, autonomous, self-acting link in the direction of coordination with the development of the system as a whole. Prices are able to fulfill this function because they are global parameters reflecting, in condensed form, information about all aspects of the economy's performance.

In general, a network can be developed by moving backward from the goal to find the ultimate criterion of development for the entire network. In moving from the goal back toward the beginning, the network's set of feasible spaces is defined.

One of the revolutionary steps in the history of human consciousness occurred at the moment when our development had reached such a stage that we could ask ourselves, "Why not begin parallel development from the start? More exactly, why not develop a network in parallel taking any arbitrary link as a starting point, assuming it impossible to construct any local criterion that would allow that link's development to be tied in a complete and consistent manner to the rest of the intercoordinated network?" This kind of parallel approach (akin to digging a canal from both ends) could reap tremendous rewards. You see, at the earliest stages of development, that is stages closer to nature, the network contains some links that produce more universal goods. If the modernization of these links could be achieved earlier, it could accelerate development of the system.

However, these tremendous gains exact a heavy toll. The problem is that at the beginning the potential benefits may not be entirely clear. This means that the immediate connection does not exist between the starting link and the rest of the world, or in terms of our preceding discussion, the stage of development individual links have achieved is not complete and consistent with the existing network. For this reason the development of the starting link easily could be fruitless or reach a dead end.

The two-fold source of development — from the beginning and from the end, coupled with the described inconsistency of the process of development of the network both from the beginning and from the end — is not the exclusive prerogative of the economic system. These features are characteristic of the creative process in general. In fact, in a number of areas of human endeavor, this two-fold source of creation was tapped much more amply than in economics.

Before I proceed to discuss the two-source origins of development drawing upon the history of the field and the relevance of the results for the analysis of the economic system, I'll try to formulate the notion of the measure of development of a given field. First, we must uncover the degree to which the development of a given field is directly subordinated to the solution of the applied problems arising in that field, that is, to movement from the end point, and the degree to which in this field the abstract building blocks and operations are defined and their combinations researched (movement initiated at the beginning).

Obviously, both approaches are more developed where the given field is more developed. In contrast, if in a field of activity only one approach has been developed, that field is still at an earlier stage.

The matter becomes considerably more complicated in finding weights of each approach that are probably indistinguishable from the long-term perspective and may hold different value at some particular stage of development. On the one hand, the respective weights depend on the extent to which a lack of fundamental individual pioneering ideas or lack of a deductive scheme, designed to conceptualize the findings in the R&D sphere, impede development. The rival factor, which also affects the assigned weights, is the amount of the available resources. But no matter how these approaches are tallied up, it is apparently acceptable to say that the most developed scientific or artistic fields are mathematics, painting, and music because both approaches are represented in them to the highest degree.[4]

On the Depth and Breadth of the "Two-Source Origins" of Development of the Economic Network

One could assume that as the network develops, the breadth of duality increases. This is supported in particular by the fact that the rates of growth of the fields of fundamental science have outstripped that of production to such an extent that a new area has emerged — engineering. Until the day when science exercises a directly decisive influence over the development of production, it will continue to be allocated too few workers and resources. Given these conditions, the field of science may be ignored except to utilize the innovations it produces as they emerge.

In this connection it may also be noted that the paradigm of the development of economics is in many ways comparable to that of biology. In biology the Darwinist model has held sway for the past century. According to this paradigm, the source of change in a system is random mutations; the main element is natural selection, which follows the principle of survival of the fittest; and existing mechanisms reinforce inherited traits.

For a long time economics made use of an analogous paradigm. It was assumed that technological and organizational change occurred spontaneously and in an isolated fashion (even a barber could obtain a spinning loom, the machine that played such an enormous role in the industrial revolution). The main control is the competitive market mechanism that winnows out the less fit innovations. The presence of organizations reinforces the results of this selection process and provides for its regeneration.

Under the influence of new conditions the paradigm in biology is changing. The development of molecular biology and the uncovering of genetic structures is allowing us to observe the internal mechanisms of change. They are connected with such discoveries as jumping genes, retroviruses, introns, and z-genes. These discoveries do not negate the mechanism of random mutation but merely restrict the realm of its applicability. At the same time it is becoming clearer that conflict and competition for survival as well as cooperation and collaboration play a role in the mechanism of natural selection. Finally, successes in genetics are letting us better understand how selected-for traits are reinforced.

Something analogous is going on in economics. It is becoming clear that a new sphere within the economic system is being recognized. It is connected with the development of science (both fundamental and applied). This does not negate the role of random inventions and minor changes. They continue to coexist with science, comprising an important part of the sphere of improvement. As far as selection from among the innovations is concerned, the forces of collaboration and cooperation are at work alongside those of conflict and competition. This is especially pronounced in the formation of all sorts of professional societies and clubs. Finally, a great variety of different types of organizations is being created to support the regeneration of the process itself.

One could assume that as this or that area of human activity develops, the depth of duality of the network decreases, that is, fewer links remain that are not closely connected to their results. Take, for example, science. At first it was developed in universities and academies. Meanwhile, in recent decades science has started to be intensively developed in firms as well. Initially in the private sector it bore mainly an applications-oriented character, but gradually there, to, it began to delve into more basic issues. This shift in the relation of companies to basic science has been dictated to a large extent by the circumstance that the number of links separating fundamental science from its application has been sharply reduced.

An extremely vivid turnaround in this respect took place during and after World War II. Take, for example, the enormous successes in technological development after World War I, the development of the automobile industry, aircraft construction, and tractor building. All this to a large extent was the result of engineering genius. Of course, in the final analysis scientific ideas were the basis for all these innovations. But too many intervening links lay between science and production for the inventors and the industrialists to establish direct connections. The creation of the atom bomb was in this respect a new phenomenon. Outstanding scientists and engineers worked together in order to produce it because, without the direct incorporation of fundamental scientific findings, it would have been impossible to solve the engineering problems.

This kind of reduction in the number of links between fundamental science and its embodiment in the form of final products has also occurred because applications have become possible on deeper levels. What is called fundamental science, including pure mathematics, belongs precisely to the research of deep layers. Therefore, because of the development of physics, biology, and artificial intelligence, a correspondence is emerging between the discoveries in the realm of pure mathematics and their applications. The development of the theory of elementary particles, of molecular biology and genetic engineering, and of cognitive processes in psychology and their connection with the newest trends in mathematics are all examples of this bridging of science and its applications.

Of course, some gaps still exist between basic science and its applications, but the quantity of extra links existing between them has been remarkably reduced. This major turning point in the history of science could be very dangerous because it compresses the lag between basic research and its application. However, even a small number of links creates problems in relating basic research to its applications. Chess players realize how difficult it is to solve Loyd's problems, although they know the solution exists and that it can be achieved in a limited number of moves — three or, at times, even two.

All these deliberations regarding the causes of gaps in the economic network lead us to the problems of survival, growth, and development.

SURVIVAL, GROWTH, AND DEVELOPMENT

The principal approach to searching for the best value of the function of change must reveal the conditions under which the search is conducted and then find the set of methods best adapted to these conditions. The counterapproach is to devise a uniform method of search, that is, a general method of solving the problem. This leads us to the following rather rough classification (Ackoff, 1981) of the dynamics of an economic system denoted by three categories (logical variables) — survival, growth, and development.

From the functional point of view, survival represents an extreme case with emphasis placed on maintaining a minimum level of vital parameters comprising the system. Growth presupposes an increase (as judged by a preset criterion) in the values of parameters already incorporated into the system. Development is concerned with creating favorable conditions for the future, in other words, with creating the potential of the system that may give birth to a new system with an even greater potential.

The functional definition of development given above has a direct connection with the previously discussed breaks in the economic network. If such breaks have not taken place and if it has been possible to link all parts of the system, the category of growth (and its extreme case, survival) could be sufficient to characterize the dynamism of an economic system. The other approaches to the definition of development also confirm the last statement.

From the structural point of view, survival is aimed at preserving the existing objects; growth indicates an increase in the number of existing objects; and development implies a variety of objects (with new ones appearing in the course of development) as well as their interrelations.

From the process-oriented point of view, survival is a search for a stationary state; growth must facilitate the creation of a complete and consistent mechanism of coordinated growth of the sought-for variables; and development implies the creation of a mechanism that supports both the creation and the resolution of incompleteness and inconsistency in the existing mechanism.

From the point of view of evolution, survival, growth, and development are all interrelated but are, at the same time (at least to an extent), independent characteristics.

Let me clarify this point. An economic system can indeed survive in a stationary state for a long time. Take, for instance, the various tribes of Africa or Australia. They have managed to survive for many hundreds of years while living in a rather stationary state. Their success can be attributed to a number of factors: their religion, which reconciled an individual with death (disappearance of the body) — both his own as well as his close ones; population growth limited by means of birth control, wars, and, especially, disease having a very significant effect on the size of the population; and the ability to generate positive emotions directly by means of some easily accessible stimulants, such as drugs. All this was accomplished in the absence of any technological changes in the means of production (in some instances, technological innovations were explicitly prohibited). Particularly interesting in this respect is the history of India, a country that for the last 2,000 years and up to the present has followed a similar course of development.

Now try altering one major component without making corresponding changes in all the others, and the system is doomed to degrade, if not perish altogether. An example of this situation is the rather recent demographic boom resulting from medical intervention involving many countries and many tribes that have thus far lived under rather stationary conditions. Moreover, an economic system that, because of a scarcity of natural resources, cannot rely on a purely extensive growth of its components to develop must resort to technological improvements and, therefore, face problems associated with development.

For this reason, long-term survival is impossible without growth, and long-term growth presupposes development. Proceeding in the backward direction, it is obvious that long-term development as well as long-term growth require survival. There is still a question of whether long-term development is necessarily accompanied by growth. This seems indeed to be the case if only because long-term development presupposes survival, and survival is impossible without growth.

The categories of survival, growth, and development being interlaced does not alleviate the problem of their respective priority in analyzing the workings of an economic system. Widespread is the view that the key objective is survival, both for the economy as a whole as well as its individual sectors (and the firm in particular). Naturally, survival is vital, because it is the prerequisite of all else. We can even look at growth and development from the standpoint of survival. Nevertheless, I am a proponent of the primacy of development with growth and survival allotted a subordinate role. One reason for this is the psychological role of the mind-set that exerts a tremendous influence on the focus of the decision makers.

Perhaps, a specific example will clarify the point. A few years ago I happened to meet some leading executives from the Clark Corporation. At the time, the firm was suffering a deep crisis and was, in fact, on the verge of bankruptcy. Under the circumstances, survival would seem to be the most pressing aim thus focusing the attention of the firm's top echelon upon eliminating all unnecessary current expenditures. However, the mind-set of the firm's leader, James Rineherd, a man who has successfully combined academic and practical thinking, was geared toward development; survival was viewed only as a prerequisite. As a result, cutbacks initiated by the firm were carefully screened so as not to cut back the firm's future. Naturally, firm leaders realize this vital provision. But it takes an unusual gift to actually implement the idea of development, especially when the firm is threatened with bankruptcy, and to make survival and growth aims subsidiary.

This raises a very pertinent question: "Why distinguish between the notions of economic growth and economic development?" One reason to do so is suggested in the above-mentioned books by Ackoff and Gharajedaghi (in collaboration with Ackoff), namely the need to differentiate between the quality of life and the standard of living. The first category corresponds with development; and the second one, with growth. While these two notions are interrelated, higher quality of life does not automatically imply growth of the standard of living — a person can sacrifice material well-being in order to improve the quality of life.

The present work focuses on the process of development. Analysis of the means of development of the economic system and of the respective methods of measuring it constitutes the core of the book. The remarks I have made at the beginning of the section regarding the multifaceted systems approach, especially its structural facet, may baffle the reader. The thrust of the book is upon all these diverse aspects of the process with respect to development.

NOTES

1. The elaboration of these definitions is largely due to Russell Ackoff and arose as a result of the collaboration between him, Jamshid Gharajedaghi, and me in the course we taught (on social pathology) in the winter of 1982 at the University of Pennsylvania.

2. Equating pathology with illness is not uncommon as demonstrated vividly by Kenneth Boulding's definition of pathology: "Pathology is the study of ill health and can easily be extended to the study of the ill health of the society — that is, conditions that are almost universally recognized to be bad" (1981, p. 17).

3. A growing number of physicists (Johnson, 1989) believe that one cannot confine the study of the physical world to just the categories of matter and energy; in their opinion, the category of information, too, should be introduced into the study of the development of the physical world.

4. In mathematics both-ended development is visible over the course of practically its entire recorded history. One of these ends was the need to solve practical problems, land surveying in particular; the other was the desire to explain the harmony of numbers as such. In Egypt the development of mathematics was driven by practical needs, such as geodesy and construction of pyramids. Greek mathematics, Euclidean and Pythagorean in particular, professed aloofness from reality and the purity of mental constructs. Pythagoreans espousing the credo that "the world is ruled by numbers" despised any kind of application. Only in the nineteenth and twentieth centuries, on a deep level of penetration into mathematical structures, did the two approaches begin to fuse together to achieve cross-fertilization, for example, with number theory, analysis, and theory of probability. Naturally, this is not to say that in mathematics there are no longer any unexplained links between these two approaches.

For a long time painting developed predominantly as a representation of the concrete. Gradually, as the artists' understanding deepened, they began increasingly to move away from concrete elements in their work. This, for example, is clearly evident in the evolution of Vincent Van Gogh's painting by comparing his *Boots with Laces* with his *The Abandoned Quarry*. It is more important for me to stress that along with the painting of concrete objects there also developed painting based on the combination of abstract elements and colors. Although great artists achieved this, for a long time it was considered subordinate. Apparently only in the twentieth century, beginning with the work of Wassily Kandinsky (1946), did abstract painting start to win its right to independence. This latter development has led to archmodern art beginning to synthesize abstract and concrete structures, to the mutual enrichment of both. For example, in a number of pictures by Michail Shemiakin, a delightful synthesis — using the motif of Russian folklore — is evident between the, to some extent, concrete structures of Hieronymus Bosch and the abstract structures of Kandinsky.

2 Indeterminism-Determinism

DEFINITION OF INDETERMINISM: FUNCTIONAL, STRUCTURAL, AND GENESIS

The general category of indeterminism is poorly comprehended. An indirect confirmation of this is the fact that there is no notion of the degree of indeterminism: usually determinism is opposed to indeterminism without any intermediate phases.

Human knowledge has evolved through many stages reflecting our understanding of the concept of measure. At some early stages, as evidenced by a number of ancient languages, man was limited to binary relations denoting extremes such as black-white and hot-cold. Later stages of human evolution witnessed the appearance of phases — measure represented by fuzzy sets corresponding to certain states; at still higher stages the notion of measure is extended over a continuum.

Systems methodology can shed some light on the subject and further our investigation of the category of degree of indeterminism. Let me preempt a more detailed discussion of the specifics of a systems approach by noting that the phenomenon of indeterminism can be viewed from two different perspectives. The underlying tenet of the first approach is that the observed system and the vehicle of observation are mutually isolated; that is, neither the observer nor the system exerts any influence upon each other. The working premise under the second approach is that the observer views the interaction of several systems or that the observer actually interacts with one or more systems. In any case, one of the participants becomes the determiner and the other the determinee.

The aspects of determinism analyzed below pertain to both the first and the second approaches. However, because of the methodology of a constructive approach, which I prefer in view of certain heuristic considerations, the implicit focus of my exposition is placed upon the second approach.

Systems methodology suggests the following viewpoints from which to define the notion of the degree of indeterminism: function, structure, process, operator, and genesis. To begin with, I shall present a very condensed description of the various aspects of the systems approach.

From the functional perspective, the degree of indeterminism of a system may be defined as the degree of influence it can exert upon development as manifest in the varying degree of uncertainty. In other words, the degree of indeterminism incorporates two factors: the impact of the action and its uncertainty. The stronger the impact upon development, the more the system is determined, and, consequently, the lesser is the degree of indeterminism. From the structural point of view, the degree of indeterminism is defined as the degree of completeness and consistency of the links of the system. In terms of process, the degree of indeterminism can be viewed as the extent of amenability of the processes taking place in the system to change. From the standpoint of an operator, the degree of indeterminism reflects the power of the operator that executes the various processes. Analysis of the degree of indeterminism from the standpoint of genesis yields a measure of impact of past indeterminism upon indeterminism in the future.

In pursuing our analysis of indeterminism, it would seem highly fruitful to bridge the above mentioned aspects of degree of indeterminism with some classical philosophical problems that arise in connection with the problem of indeterminism. Table 2.1 illustrates the relationship between the two. Let me elaborate on this table starting with the functional approach.

It is widely held that determinism entails the prediction of all future (as well as past) states of the system. This particular feature of determinism formed the groundwork for the classical treatise of the problem by Pierre Laplace. Laplace postulated that given the mass, the coordinates, and the impulse of all the material particles in the universe at a given moment of time, it is possible, at least in theory, to compute all the future as well as all past events by means of Newtonian mechanics.

This brand of determinism was developed significantly in the works of an outstanding philosopher of our century, Karl Popper (1982). Popper's proposed definition of determinism presumes that: "the doctrine that the structure of the world is such that any event can be rationally predicted, with any desired degree of precision, if we are given a sufficiently precise description of past events, together with all the laws of nature" (pp. 1–2). Without going into some highly interesting particulars introduced by Popper to the concept of determinism, I merely wish to note that his concept adheres to the classical Laplassian view of the world in that it embodies the issue of predictability.

TABLE 2.1
Relation between Various Aspects of Systems Approach and the Problematique of Indeterminism

Aspects of Systems Approach	*Characterization of the Degree of Indeterminism from the Viewpoint of a Given Aspect*	*Classical Problem Formulation Associated with a Given Aspect of the Systems Approach*	
		Passive	*Active*
Functional	Overcoming uncertainty of impact upon system's development	Predictability Fate Uncertainty	The plausibility of knowing the future now Can one's destiny be changed? Is it possible to stipulate the dynamics of system's development with a given degree of certainty?
Structural	Completeness and consistency of system's network	Connectivity Openness	Can the universe be split into independent parts? Can the system withstand changes resulting from external perturbations?
Process	System's capacity to change	Reactive and selective representation of a system Free will	Is an analytical representation of a system deterministic and an optimal one indeterministic? Is the system responsible for its actions?
Operator	Power of the operators implementing the processes taking place in the system	Epistemological vs. metaphysical determinsm	Can indeterminism be overcome assuming the existence of an operator possessing unlimited power?
Genesis	The impact of indeterminism in the past upon indeterminism in the future (and vice versa)	Causality	The possibility of creating something from nothing

In my opinion, predictability (independent of the temporal sign) is not a sufficient condition for determinism. The predictability aspect emphasizes the degree of certainty of the knowledge about the future (past) states of the system. However, from the standpoint of the functional definition of determinism given above, the central point is the ability to influence development. Perhaps, this statement will become less obscure if we consider another classical problem that springs from the notion of determinism, namely the problem of fate. Those who believe in destiny may purport that the manifestations of fate are absolutely elusive for the human beings, and yet fate is so powerful as to determine the course of development of all objects under its control.

The last part of the statement about the omnipotence of fate may well raise the following objection: "Does not this view of the relation between fate and indeterminism contradict the above definition of indeterminism based on two components — the extent of influence upon development and the degree of uncertainty?" This question is indeed justified; hence, it can be reasonably assumed that the states fate leads us to might be completely uncertain. Naturally, man attempts to discover these states through astrologics, seers, and, at times, even resorting to science, which essays to establish the laws of development, but the effort to discover the future may still prove futile.

At the same time, acknowledging the absolute power of fate, that is its inevitability, as the expression of complete determinism does not contradict the idea of fate manifesting itself with complete certainty. The crucial point is the set of parameters over which this certainty is perceived. These parameters may assume the form of states of a system or directed action of its operators. Indeed, as far as knowledge of the states of the system, conceding the inexorability of fate may imply complete uncertainty. A statement to the contrary would predicate connection between the inexorability of fate and uncertainty with regard to the actions performed by the operators of the system. It is precisely this unequivocal knowledge that the aimed actions of the operators bear no impact upon the development of the system that corroborates the idea of fate's inevitability. It seems that the lesser the knowledge of the relative influence of the operators upon the development of the system, the smaller (other conditions being equal) is the determining role of the system.

Therefore, predictability merely measures the degree of certainty of one's knowledge about the future (past) states of the system, which, in turn, is vital in uncovering the capacity of the determining object to affect development. Other conditions being equal, information about the future helps anticipate future events, make the necessary preparations, and perhaps even alter them. At the same time, the stronger the impact exerted by the determining object upon development — as judged by the information available at a given moment — the less knowledge about the future it requires in order to channel development along the desired course.

The distinction between the degree of knowledge concerning the future allows us to discriminate between unexpected outcomes and mistakes. Indeed, the two notions are quite different. A mistake presumes the method of development must be known to the decision maker. But because he does not know it or forgets it, his actions, as a rule, lead to negative consequences. An unexpected outcome means that the decision maker has no way of knowing the consequences because there was no method for determining the best course of action. By dealing with unexpected outcomes, the decision maker expands his knowledge and ability to create more sophisticated methods of action in the future.

A number of specifically pragmatic consequences ensue from this basic difference between the notions of unforeseen outcome and mistake. For instance, what is the proper working environment for the innovators? In case the innovators' decisions yield to verification, what is needed is a thorough system of expert evaluation in conjunction with high rewards and severe punishments depending on the actual results of implementation. However, the situation where the ensuing consequences are not amenable to earlier verification calls for a different organizational structure based either on the financial independence of the innovator or, in case of corporate ownership, upon trust for the innovator, within certain limits, combined with a system of high rewards and weak punishment depending on the results of implementation.[1] Hence, the degree of certainty of knowledge about the future (as well as past and present) looms large in establishing the degree of determinism.

The next most popular, if not more popular issue (following determinism/predictability and determinism/fate), is the study of the relation between determinism and causality. The *Encyclopedia of Philosophy* (1972) presents the following definition of determinism: "Determinism is the general philosophical thesis which states that for everything that ever happens there are conditions such that, given them, nothing else could happen."

This definition reduces determinism to the problem of causality. This reduction reminds me of the age-old debate in economics concerning the relation between costs and benefits. According to classical political economy as expounded by Smith, Ricardo, and Marx, the value of the output is measured by the expenditures of one particular factor employed in its production, namely labor. The advent of the utility theory drastically altered the concept of the relation between costs and benefits. The latter were assigned autonomous value through the utility of the consumers of these benefits. Moreover, all the factors involved in producing the benefits were evaluated based on their contribution to the volume of benefits produced.

Perhaps, in spite of its arbitrariness, the above analogy reveals that causality and impact over development represent two facets of the problem of determinism. Reducing the problem to one particular aspect of causality serves to obscure the issue, if only by denying explicit emphasis upon the outcome.

In fact, the presence of indeterminism was frequently explained solely by a lack of sufficient causes determining the outcome. This approach raises a number of questions: Are the causes defining a particular outcome lacking in principle? Are existing causes difficult to uncover? While the causes do exist, are they undiscoverable in principle at each given moment of time?

The first question effectively implies that every change has its causes; that is, there are always causes that induce a given change. This view stems from a rather simplistic notion that cannot conceive of a change without a cause, just as we cannot imagine something originating out of nothing. Apparently, striving for a greater degree of determinism (in a sense of greater influence over future development), as expressed in the respective degree of uncertainty, entails finding new causes that exert stronger impact or structuring the existing set of causes in a more coherent fashion. Naturally, the set of causes affecting a given system is limited, so not every system can be fully determined. Still, the fact that an event has taken place or can take place means that a sufficient set of event-originating causes exists.

Of course, it is one thing to affirm the existence of sufficient causes exerting some influence upon development and quite another to expose these causes. As the second question points out, these causes may be extremely difficult to discover. In the discussion of the problem of uncertainty, I tackle this problem of uncovering causes that stipulate deterministic development.

The last question concerning the relation between causality and determinism considers a situation when determinism-inducing causes do exist but are theoretically undiscoverable at each given moment of time. I address this issue in the light of the effect in quantum mechanics.

The assumption that control over development with a degree of certainty presumes the existence of event-originating causes establishes a close connection between the principle of causality and the principle of determinism. However, I would think it detrimental to reduce one principle to another because each one emphasizes different aspects of the problem of development.

Table 2.1 describes the relation between the various aspects of the systems approach and the issues of indeterminism. The structural facet of the systems approach to the degree of indeterminism gives

birth to such notions as connectivity and openness of a system. As far as connectivity is concerned, let us interpret development as a creative process where some parts are created in parallel and are thus not interlinked with each other. Viewed from this perspective, any unfinished creative venture may incorporate unconnected parts. Trying to establish the degree of indeterminism, as long as this is based upon measuring the holistic effect, presupposes measuring the innerconnectivity of the various parts. Other conditions being equal, the weaker the links among the various parts of the system, the weaker and the more uncertain is the influence they exert upon development.

The other structural factor affecting the degree of indeterminism, the system's openness, poses much greater problems. It seems obvious that an open system implies indeterministic changes because the external factors affecting its development are unaccountable. However, things are not so simple. In principle, it is possible to furnish specific conditions under which an open system cannot be determined by the outside world. This requires that a given open system possess much greater powers than the external world or that it have a protective shield that would repel all the disturbances coming from the environment or that it possess sufficient flexibility (and adequate reserves) to assimilate all the perturbations flowing in from the environment.[2]

Of course, the above conditions are rarely fulfilled, so an open system is primarily indeterministic. In a nutshell, the problem of determinism in a closed system can be stated in the following question: "In a closed system can a demiurge determine development completely?" It seems to me the answer to this question is negative. To corroborate my position I shall examine the process aspect of determinism.

DEFINITION OF INDETERMINISM: PROCESS AND OPERATOR

The process aspect of determinism is examined in the light of its connection to the general systems issues and to the classical problems of philosophy.

The major problem raised by relevant general systems considerations is the mode of representation of a system. I distinguish two broad categories of methods of representation: reactive and selective. The reactive class presumes the existence of a law linking a system's inputs and outputs. Here, inputs play the crucial role, because they determine the output via the governing law. This mode of representation creates an impression that the system is completely determined. The selective class of methods presupposes certain degrees of freedom brought about by the forces of attraction and repulsion. The balancing of these forces carried out by means of an algorithm constitutes the core of the system's operation. Under the selective mode of representation, the system appears to be indeterministic.

Under certain conditions, the reactive and the selective modes of representations turn out to be isomorphic. However, many researchers regard one way of representation as deterministic and the other as indeterministic. Both of these modes of representation can be completely deterministic however, because an algorithm fulfills the same determining role as a law. Both an algorithm and a law are merely two ways to represent programs that spell out the behavior of a system.

The problem of program hierarchy is another dimension of the process aspect of indeterminism. The category of a program as an object of investigation is a rather latecomer to the scientific scene, and it owes much to the strides made by twentieth-century science. The invention of computers, the study of genetic code at the molecular level, the technology of mass production of complex goods — all these advancements represent areas of human endeavor that call for rigorous (sometimes even formalized) determination of complete and consistent connections among the various links of the mechanism. In other words, they called for the introduction of the notion of a program.

A complete program implies that the system that operates based on that program is potentially as well as effectively deterministic provided it ensures 100 percent warranty of the resulting output. Now, in keeping with the functional definition of indeterminism given at the beginning of this chapter, a system based on a complete program whose output is not completely determined — if only because the inputs of the system are characterized by uncertainty — will be regarded as being indeterministic.

A system based upon a program that is incomplete, but fixed (incapable of changing), is deemed potentially indeterministic. In case the output turns out to be fully determined, the same system could probably be regarded as actually deterministic. The capacity of a given program to change reflects actual indeterminism. One particular case, when changes taking place in the programs as well as the resulting outputs of the system can be completely determined, represents another instance of an actually deterministic system.

It can be reasonably assumed that, in the more general case, enhancing the programs' completeness cannot, in principle, guarantee full certainty. This statement paraphrases my previous conjecture regarding the impossibility of a perfect foresight even for closed systems.

Let me analyze this point within the framework of the scheme of a hierarchy of programs. A behavioral program, which, according to Herbert Simon, governs the behavior of the system directly, will be denoted as a zero-level program. A first-level program changes the zero-level program. Changes in the zero-level program can occur as a result of new information gathered in the course of a program's implementation as well as latent possibilities inherent in the first-level program. A second-level program can change the first-level program as a result of new information discovered in the course of the latter's implementation as well as possibilities embodied in the second-level program itself. Moreover, assuming isomorphism among programs of various levels, we cannot discount the possibility that the first-level program can, in turn, have feedback and affect the second-level program. For instance, a third-level program can induce changes in the second-level program (because of new information accumulated over the course of execution or because of the possibilities intrinsic to the third-level program itself) and can itself be altered because of changes in the second-level program.

The described hierarchy of programs resembles the system of education. The affinity is revealed if we consider that, as a result of learning (based upon one's own experience and deliberations as well as the experience and deliberations of others), a given-level program causes changes in adjacent-level programs. However, the real focus of this hierarchy is the opportunities of programs at a certain depth to modify each other independently of the results of interaction of a given-level program and its environment.

Foreseeing the future entails the creation of a preventive mechanism. This mechanism is constructed upon certain substratum that permits simulation of future states faster than they actually arrive. However, the capacity to simulate feasible events is limited due to the volume of information and the processing speed. Therefore the process of simulation takes time. Furthermore, the preventive mechanism may incorporate a hierarchy of programs similar to the one described above. Suppose even the first-level program is incomplete and requires further improvement. For now, let us introduce one simplifying assumption that the process of development advances based solely on experience acquired during the course of simulation (as a rule, experience of the system's real-life performance is essential as well).

Current behavior could change if knowledge about the distant future (of course, distant enough so that future states bear an impact upon actions in the present) was available now. However, in view of the limited capacity of the mechanism of simulation, it takes time to discover these remote outcomes as well as to revise the program based on the new data. Further inclusion of higher-level programs into this process (keeping the above assumptions of a lag in the resulting information), exacerbated by the fact that the greater the depth of a given program the more profound the changes it can effect in programs closer to the surface of a given system, can make the search for future states infinite. However, even assuming that this process is finite, the time expended to simulate the future may be so excessive (coupled with the actual duration of the processes taking place in the system) as to prohibit the implementation of respective improvements in time required to embed simulation-suggested improvements into the initial system. At the same time (keeping the above assumptions intact) in dealing with a developing system, the following situation might arise. A system must take some step now. The outcome of its current action really stems from only a partial picture of the future, which the mechanism of simulation has managed to construct up to the time the step is taken. At the next time frame, this mechanism might advance in foreseeing a more distant future, and the newly generated image of the future might suggest certain subsequent actions other than the ones suggested at the previous iteration.

Consequently, making a series of assumptions that the current actions affect the future when simulation of the future takes place, that the information about the future, in turn, affects the present, and that the mechanism of simulation carries out all these processes with a certain lag (at the overall speed lower than the actual duration of the processes taking place in the system) guarantees that even in a closed system no deity has the ability to completely channel the future along the desired path. The full force of this statement becomes evident if we take into account the fact that improving the program requires not only the information generated by simulation but also knowledge of the system's real-life (if only because an exhaustive description of a system is impossible).

The idea of program hierarchy is also beneficial in clarifying the process aspect of the problem of indeterminism as it pertains to society. Some issues raised in this connection are free will and, from the more constructive perspective, the problem of personal responsibility.

Mere capacity of a first-level program to alter a zero-level program is sufficient to speak of the zero-level program as having some degrees of freedom. In fact, the notion of will can be understood as the ability of the system to induce its first-level program to redirect the zero-level program along a given path.

Apparently, humans possess a program (first-level program) that they employ to structure their behavior (zero-level program). Generally speaking, by utilizing the first-level program, individuals can rehash the problems posed before them and find appropriate solutions based on the observed changes. These changes might reflect not only the inner changes in their mind (in this particular case, the first-level program), but also more frequently changes in the environment. Thus, a first-level program allows individuals to

assimilate information coming from the environment and reformulate tasks based on that information. Free will seems to be manifest primarily in the ability to reformulate the problems faced by taking into account changes in the environment.

The problem of responsibility illustrates the above discussion. One aspect of responsibility is the fact that the environment imposes certain restrictions expressed in such social values as thou shall not murder and thou shall not steal. An individual may possess other values and enjoy killing others, but his liability is determined by his capacity to perceive the requirements imposed by the environment and consequently adjust his behavior. If an individual is incapable of such acts, he is not held responsible. Indeed, insane people are regarded as not being responsible for their actions precisely because their program of behavior fails to fulfill all or just one of the necessary conditions: it fails to assimilate conditions instituted by the environment; it fails to bring into accord the values of the first-level program; and finally it fails to generate action designed to implement the new program.

A criminal is deemed a criminal in a given society because he is considered capable, at least in principle, of altering his conduct according to the dictums of the environment but failed to do so because his individual system of values accorded him greater satisfaction.[3]

The final aspect of the systems approach to indeterminism is the power of the operator that executes the process. Assuming programs can ensure completely deterministic changes of the system; that is, metaphysically the system is determined, then the actual implementation of this possibility hinges on the power and reliability of the operator. In view of the limitations on the power of the operator, the problem should be rephrased to focus not just upon the theoretical existence of a program that ensures deterministic changes, but also upon a program that can be handled by the available operators. In case such a program does not exist, we are facing an epistemological impossibility of determinism.

A simple example should illustrate the idea. The game of chess constitutes a system where the number of possible combinations of the pieces (under current rules) is finite. Therefore, there is an algorithm that could reveal the optimal strategy, that is, which side determines the behavior of the other — guarantees victory (of course, there could be another type of deterministic outcome, a draw). The most primitive but conclusive algorithm would be a brute-force, exhaustive search of all possible combinations. Because the number of such combinations is approximately 10^{120}, the execution of such a simple algorithm, even by the fastest operators in the guise of computers operating at the speed of light, might take hundreds of thousands of years.

A deeper probe into the operator problem leads us to another question: "What is the impact of operator-specific features upon the degree of objectivity of an operator's determination of the system?" The notion of the degree of indeterminism immediately raises the problem of the extent to which this degree can be evaluated objectively. In the operational context, the term objectively signifies a method that, if mastered by any operator, independently of the operator's own particular features, will produce the same results if given the same initial information.

Determinism viewed from this perspective presents another extreme case — complete subjectivity. It appears in cases where the truth or falsity of the consequences of action cannot be proved. As a result the specific nature of a given operator, that is, its subjectivity, determines the evaluation.

Housed between these two extremes are states combining subjective and objective elements. The question of the interplay of subjective and objective factors and operator-specific features will be clarified when I explore the method of potentiating from the standpoint of the systems approach.

Having given a brief overview of one possible approach to the problem of indeterminism, I now focus on the degree of indeterminism.

PHASES IN THE UNCERTAINTY SPECTRUM

Generally speaking, the evaluation of the influence exerted upon development and the uncertainty of the exerted influence are closely related. For instance, depending on the initial information about the process and the operators that execute it, measuring the impact on development can simultaneously reveal the probability of this particular impact.

However, my investigation of indeterminism spotlights the problem of uncertainty because it is more elaborated than the problem of indeterminism in general. The advancements made in this branch of knowledge are relatively new.

The Aristotelian concept of the degree of certainty of our knowledge distinguishes just two possible states: certainty and uncertainty. Introduction of probability as a measure of certainty appeared much later in the seventeenth century following the work of Blaise Pascal, a great French religious thinker, writer, mathematician, and physicist. Only in the twentieth century, because of the work of an outstanding Soviet mathematician, Andrey Kolmogorov, did the theory of probability acquire a deep mathematical foundation linking it with the theory of measure.

However, many dark spots remain in the problem of finding the degree of uncertainty. A broader view of the concept of uncertainty

reveals its relevance to the qualitative as well as quantitative description of objects.

Quantitative uncertainty is commonly represented (Luce & Raiffa, 1957) as a threefold transition from certainty to uncertainty with probability squeezed in between.

> As to the certainty-risk-uncertainty classification, let us suppose that a choice must be made between two actions. We shall say that we are in the realm of decision making under:
>
> (a) *Certainty* if each action is known to lead invariably to a specific outcome (the words prospect, stimulus, alternative, etc., are also used).
>
> (b) *Risk* if each action leads to one of a set of possible specific outcomes, each outcome occurring with a known probability. The probabilities are assumed to be known to the decision maker. For example, an action might lead to this risky outcome: a reward of $10 if a "fair" coin comes up heads, and a loss of $5 if it comes up tails. Of course, certainty is a degenerate case of risk where the probabilities are 0 and 1.
>
> (c) *Uncertainty* if either action or both has as its consequence a set of possible specific outcomes, but where the probabilities of these outcomes are completely unknown or are not even meaningful (p. 13).

The qualitative measure of uncertainty concerns the degree to which the qualitative characteristics of the system's elements are known. At the limit, this could be expressed as the complete knowledge of a system's qualitative parameters or the total ignorance of them. In a relatively simple case, this degree of qualitative uncertainty is exposed as the object is being identified (as pattern recognition is executed). The working assumption here is that the object being discovered (uncovered) has already been created. An example of this kind of qualitative uncertainty is a die with all the usual numbers. However, it may be difficult to identify the observed pattern on the die if the depicted numbers are vaguely drawn. The pattern's degree of qualitative uncertainty can be uncovered using the techniques of probability (Domotor, 1976).

Various cross-combinations between qualitative and quantitative uncertainty could be generated. Varying degrees of uncertainty correspond to different methods of a system's operation. In dealing with certainty and probabilities situations, the future can be programmed by means of extrapolation or optimality techniques. The latter, in turn, embraces two types of approaches. The first approach entails designing a plan-schedule ensuring the certainty of the outcome by having the necessary reserves (for example, a public transportation schedule). The second approach advocates the creation of plan-instruction, or so-called contingency planning, that is, planning based on unforeseen events (an example of this is planning in a fire station).

Devising methods of a system's operation under the conditions of uncertainty is much more complicated. Two basic approaches to this problem have been charted. The first one attempts to reduce the problem of uncertainty to probability, and the second calls for the creation of flexible structures.

The first approach is the more elaborate of the two. One way to implement it is to simplify or modify a given problem, somehow reducing it to a class of known problems that yield to probabilistic solutions based on the statistical information about their performance. This very attractive but tricky approach has steered many scholars off track.[4]

The drawback of this approach is the difficult but very crucial problem of discerning the importance of dissimilarities between a problem at hand and the set of previously scrutinized problems. In fact, the complexity of measuring the importance of these dissimilarities has produced major methodological problems calling for new more sophisticated solutions.

These difficulties were brought to light by Frank Knight (1971). He introduced a taxonomy of probabilities that, besides distinguishing between a priori and statistical probability, introduced the so-called estimates. The latter class is described as:

> Estimates. The distinction here is that there is *no valid basis of any kind* for classifying instances. This form of probability is involved in the greatest logical difficulties of all, and no very satisfactory discussion of it can be given, but its distinction from the other types must be emphasized and some of its complicated relations indicated (p. 225).

In case uncertainty cannot be reduced to probability by means of the above method, three alternative methods have been suggested: rather than observe highly uncertain behavior of individual objects, we look at the behavior of the total ensemble formed by these objects because the latter may be more amenable to statistical patterns; accumulation of information by trial and error; and assuming some a priori probability distribution and then adjusting it (thus deriving an a posteriori probability distribution) to conform to the acquired data. Various considerations enter into the choice of the initial a priori probability distribution including the feasibility of the worse outcomes.

Naturally, the scope of applicability of the above methods is limited. For instance, under the third method, effective computation of probabilities sufficiently descriptive of the real process is frequently precluded, if only because expenditures of time and resources required to teach the system can be prohibitively high.

The second approach, designed to ensure smooth operation of a system under uncertainty, entails the creation of flexible structures. (Sometimes these structures are expressed under the terms vulnerability, resilience, and robustness.) The primary factor contributing to overall flexibility of these structures is the flexibility of the factors of production, including versatile equipment, materials, and labor, as well as a handy variety of reserves. Meanwhile, a flexible system may require not only factor flexibility, but also a adequate network linking it to other systems.

The above methods operate under the assumption that there exist algorithms linking all the inputs and outputs of the system in a complete, consistent, and optimal manner even though the information regarding the inputs and outputs is uncertain.[5] This is a very strong assumption because such algorithms, assuming they exist, may be difficult to construct or implement.

Besides, the above mentioned drawbacks of the flexibility concept suggest its reformulation within the framework of a more general concept of adaptation. This concept, understood in the broad sense of the word, encompasses not only the system's ability to respond passively to a turbulent environment, but also to actively induce its environment, channeling unforeseen turbulences along a favorable course, and alleviating the negative effects of these perturbations.

Measuring this kind of potential (which, if we take the general case, is reasonably considered unique) cannot be reduced to probabilistic methods. At the same time, it is quite plausible that complete uncertainty can be avoided and a search for orderly methods of evaluation found.

The considerations on the limited applicability of program-like methods of solution to the conditions of uncertainty and the need to introduce a more sophisticated definition of an adaptive system have led me to conclude that, perhaps, the uncertainty spectrum, which includes such well-established polar phases as certainty and probability at one end and uncertainty at the other, should be broadened. Eventually, as these phases evolve they may transform into phases that yield to probabilistic methods, just as probabilities transmute into certainty with the accumulation of sufficient information. The uncertainty spectrum can be represented in the following fashion. First on the list is the phase of complete uncertainty characterized by a lack of stable rules of interaction between the constituent elements and any kind of regularities. This phase merits special consideration because it marks a profound break between the methods of operation corresponding to different degrees of completeness of information. This lack of regularities within the phase of complete uncertainty predicates the creation of mechanisms designed to expand the manifold of the global avenues of development (this becomes the key objective) while the task of bringing the system to an equilibrium plays a subservient role. The paramount aim of the subsequent phases is the elaboration of methods facilitating movement along the selected path. Even if some of these methods assign primary importance to expanding the manifold (with balancing relegated to supporting role), this task will be executed within the confines of the previously selected course of development.

Complete uncertainty is superseded by a phase called chaos (Gleick, 1987). The theory of chaos pinpoints certain strange attractors, that is, conditions exhibiting certain stationary features, governing the dynamics of the system. A number of constants reflecting the structure of the process, the so-called Feigenbaum numbers, have also been discovered. Thus, chaos revealed a number of patterns, and these patterns introduce certain objective globalization to this dynamic process. However, it seems to me that the discovered patterns are not universal; that is, they encompass only those cases of uncertainty that reveal some stability of the rules of interaction among the elements of the system.

The phase of chaos is followed by the phase of predispositions, and following that is the phase of completions. For the purposes of this exposition, I focus on the latter two.

The methods of operation employed at the phases of certainty and probability were based on the assumption that the chain (or part of the chain) of transformation of inputs into outputs can be linked in its entirety thus making it possible to forecast the future or, at least, to ensure that the selected methods of operation will guide the system along the optimal course of development in case of exogenous perturbations. In other words, the possibility of constructing inner-complete solutions was deemed plausible at both the certainty and the probability phases. Subsequent improvement of the methods employed at these phases would not alter the final results, a condition that spotlights the affinity between the certainty and probability phases and the phase at which the completions are formed.

Not all problems, however, yield to solutions in terms of completions. Herbert Simon (1957) initiated a rather thorough analysis of this class of unyielding problems, elaborating the so-called satisficing approach.

As fruitful as the concept of satisficing is, in my opinion it has failed to provide a constructive principle, which would make it possible to solve problems under certain conditions, characterized as being between chaos and completions. This point was brought to the reader's attention when I stressed the need to create potential in adaptive systems. In this context, potential signifies a certain state that improves the system's position — a position that, on the one hand, certainly affects development but without knowing the probability of the influence it exerts but, on the other hand, purports to be more certain than chaos. My aspirations have to do with the development of certain principles underlying the creation of such potentials. The impact exerted by the potential upon future development, as manifest in the respective degree of uncertainty, is called the degree of predisposition. The method of computing this potential is called the calculus of predispositions. The core of the book is devoted to elaborating this method.

INDETERMINISM IN A HIERARCHY

My concluding remarks to this chapter concern the various levels of systems. The above considerations regarding the category of indeterminism pertain to systems at the same level or to subordinated and commanding systems. But they apply equally well to hierarchically organized systems, that is, structures shaping systems that incorporate other systems as subsystems or that are part of a larger suprasystem.

An economy as a system is open in relation to its surroundings — suprasystem, which included both nature as well as the sociopolitical system; at the same time, an economy is constituted by a group of subsystems: branches, firms, and regions. The correlation between subsystem and system, generally speaking, could assume isomorphic interaction between the system and suprasystem. To simplify matters, the measure of degree of indeterminism in a hierarchy will be based on a binary representation, namely determinism and indeterminism. Figure 2.1 presents a matrix to establish the connection between the above mentioned various phenomena.

FIGURE 2.1
Matrix of Interrelations of Degree of Indeterminacy and a System's Structures

Indeterminism	Determinism	
	Suprasystem	System
Suprasystem		
System		

From the operational viewpoint, this classification scheme means nothing other than a division of system representation into two stages: formulation and solution. The suprasystem dictates the mode of problem formulation; the system's structure, its method of solution. If we bear in mind that the suprasystem, in turn, is also a system in relation to its suprasystem, and a subsystem of this system is itself a system with its own subsystems, then division of the problem into two stages becomes relative. What is problem formulation for one system is a stage of problem solution for its subsystem, just as problem solution for this system includes problem formulation for its subsystems. If we consider that the world is infinitely broad and deep, then the aforementioned relativity of problem formulation and solution becomes absolute.

The relativity of the problem formulation stage and the solution stage are most obvious in analyzing determined suprasystems that include determined systems. Examples of this situation are easy to see in decomposition linear programming procedures used to solve large-scale optimization problems.

Combinations in which both structures, the system and subsystem, are undetermined are relatively simple to analyze. More complicated are cases where one of them is determined and the other is undetermined. This kind of situation may arise under the following conditions. Say the system is deterministic and the suprasystem indeterministic. This would mean that internally the system's links between its subsystems would be organized on the basis of a complete and consistent program. However, there is no such program in system interactions in the frame of the suprasystem. Therefore, the suprasystem will require an indeterministic representation of the system in parts that relate to external demands. The system's internal constraints in this case will be given deterministically. This is caused by the fact that within the system the representation of individual subsystems in external parts will be given deterministically, that is, the form of each external part of the subsystem representation will be invariant relative to the form of suprasystem representation. The form of representation of subsystem internal constraints will depend on the deterministic or indeterministic character of subsystem performance.

The final situation has the suprasystem determined and the system undetermined. In this instance the system will be represented externally as a deterministic one, but its internal constraints will be given in an indeterministic form.

A subclass of system and subsystem interaction may be distinguished when the subsystems are indeterministic and the system as a whole can represent its internal constraints deterministically. This kind of situation occurs when the number of subsystems is so great and the forces guiding them so varied that as a whole their behavior presents us with a stable statistical ensemble. By assuming stability in the averaged parameters of this subsystem aggregate, we are able to represent the system's internal constraints deterministically. Situations resembling the described ones are not limited to physics. They also belong in the realm of economics — provided economics is regarded as a suprasystem, the totality of consumers as a system and individual consumers as its subsystems.

NOTES

1. This approach sheds new light on the story of the flood as told in the Torah, that is, on the old question, "Can the Creator endowed with the powers to foresee everything destroy His own creations?" The standard explanation for the destruction of Man by the flood states that people, having free will, chose to act in a way unacceptable to God and were punished accordingly. But animals and plants also suffered destruction and the earth was flooded with water. These objects had no free will, did not wallow in sin, and could have been put by the Creator to other purposes. One is hard pressed to find a rational explanation for God's actions if one sticks to a deterministic framework. Indeed, why would an omniscient and omnipotent God with the power to foresee the future completely destroy his own creation not sparing even things that lacked free will. An idea of God as a developing entity acting within an indeterministic frame will perhaps yield a rational explanation for the flood: it could be viewed as God's response to unforeseen outcomes. At the same time, God's great might and knowledge are manifest within an indeterministic scheme in God's ability to respond to these unforeseen outcomes and channel the events along the desired course. God's actions can thus be interpreted as eliminating undesirable, unexpected outcomes arising in the course of the development of the universe rather than as God correcting his mistakes. (See also Katsenelinboigen, 1989.)

2. A country's military defense mechanisms fit into this scheme. A country may possess a military force capable of destroying any attacking force; or it may be surrounded by impenetrable walls; finally, the country may be able to assimilate the aggressors within its own culture.

3. Punishment for the crime committed does not make any sense if the individual is incapable of changing his behavior. More precisely, an individual held to be a criminal must merely be isolated from the environment because he poses danger to it. All attempts to remodel a person by means of punishment should in this case be deemed futile. By predicating free will, however, the administration of punishment, that is, making more severe in the mind of the individual the negative judgment assigned by society to certain acts, one can try to modify the person's behavior, in other words, achieve a change in his behavior *post factum*. The other implication of the idea of free will is that punishing certain individuals helps inculcate in the minds of others the knowledge of the negative judgment pronounced by society regarding certain types of behavior, which represents an attempt at a preventive approach to tempering the behavior of people.

4. I believe Herbert Simon fell victim to this temptation when, in the 1950s, he proposed to create a computer chess algorithm superior to the world champion in a ten-year period. In no way do I mean to downgrade the wonderful scholarly achievements of Simon and his colleagues in this area. What is important is not to reject their approach completely or embrace it with no reservation — or subject their methods to the extreme test of solving the problem in its entirety — but to find a proper place for their discoveries in the overall scheme.

3 Degree of Indeterminism and the Performance of an Economic System

SYTROPY — A FUNCTION MEASURING THE DEGREE OF INDETERMINISM

Degree of indeterminism as a philosophical category implies the existence of a criterion associated with that category. This criterion is the two-faceted criterion of development that incorporates two factors: the extent of influence exerted upon development and the uncertainty of the exerted influence. One particular (degenerate) case of the criterion of development is the criterion of growth.

As is well known, the most general characteristic of the physical picture of the universe, one reflecting its dynamics, is entropy, that is, the degree of disorder in the system (its counterpart is the notion of negentropy, the degree of order). Closed systems, and physical systems represented as such, tend to greater entropy. Open systems evolve differently tending to greater negentropy.[1] As a result, this category won citizenship rights in such open systems as biology (Bertalanffy, 1968) and economy (Georgescu-Roegen, 1971).

However, systems interacting with a given open system, as well as the suprasystem, may exhibit an increasing degree of entropy. Let us examine two systems adjoining the economic system from this perspective. One is the ecological environment; the other is human psychology.

In principle, development of a given system can inflict damage to other systems with which it interacts. This applies to the alliance between the economic system and nature. Ever-increasing negentropy of the economic system entails the depletion of natural resources, pollution of the environment, and violation of the ecological balance causing growing entropy in the environment. Excessive depletion of resources and environmental pollution can eventually result in the growing entropy of the suprasystem, which includes the economic system. We must exercise extreme caution in

setting the pace of negentropy growth for a given system and try to ensure parallel growth of negentropy in related systems.

The connection between the psychological realm of man and the economic system is manifest in man's deciding upon the general course of economic development. The utility function formed by man's psychological system determines personal values assigned to different things and manifests formally the guiding forces of a given person.

Viewing the economy as a closed system, man's drive to optimize his utility function can be regarded as a strife for greater entropy (Pikler, 1955). Here, man's behavior is very similar to the behavior of a physical system.[2] We can also treat the utility function as aspiring toward greater negentropy if human activity is assumed to be taking place in an open system.

Negentropy of the processes that generate the utility function within the human psyche may also increase: as our body of knowledge expands, the function becomes more and more ordered. However, greater order is laden with dangers. For instance, if the utility function forms along extremely extrovert lines placing a high value on the fulfillment of today's wants and assessing the destruction of the environment as insignificant, it can lead to the growth of entropy in the suprasystem and the ensuing consequences for the economic system itself.

The category of entropy is paramount but insufficient to characterize the process of development in its entirety. Variety of objects, that is, heterogeneity of a system, is another category useful in representing development.

Studying the process of change in the socioeconomic systems, Jamshid Gharajedaghi (1985) viewed these changes as stemming from diversity and order that correspond to differentiation and integration respectively. He treated these two variables as a two-dimensional entity rather than a dichotomy in a one-dimensional structure. Each of these parameters can assume a range of values. Thus, development is represented as a function of two independent variables: differentiation and integration.

Associating greater differentiation with an increase in complexity and greater integration with an increase in order, Gharajedaghi (1985) notes with respect to the living systems: "that movement toward complexity and order is the essence of the negentropic processes in living systems" (p. 40).

Gharajedaghi's broader definition of entropy might not be warranted. We can keep the old definition of entropy as the degree of order and invent a new term for the development function. I have termed this function "sytropy," for it reflects on systems turn. Of course, terminology is not the issue here; the issue is measuring the degree of development defined by Gharajedaghi.

Gharajedaghi's concept of entropy is limited to living systems. Physical systems, in his opinion, evolve in a one-dimensional fashion toward greater complexity of the structure of matter whereas biological evolution reflects movement toward greater complexity as well as greater order (1984, p. 37). I have pointed out in the previous chapter, physical systems, if viewed as being self-perfecting systems, can also be viewed as a two-faceted process directed at both complexity and order.

Thus, the dynamics of an economic system is assumed to unfold along the said dimensions that interact but, at the same time, are both independent. Indeed, let us take one extreme case where variety is actually diminishing while the degree of organization rises. This sameness of all people and the goods they consume (especially cultural ones) — being a prerequisite of order — represents the dream of many totalitarian leaders striving for maximum order. Such a system, however, will not flourish. Eventually, it is doomed to stagnate even if at some point it does show an increase in production. Another extreme is primitive anarchy, which proclaims uninhibited growth of variety and completely rejects any methods designed to coordinate human activities. Ultimately, the two systems may reach the same state. Under anarchy, the strong ones might seize power, suppress the weak ones, and impose their own uniform rule upon all the people.

It seems the term "progressive country" refers to countries where the growth of variety is accompanied by increasing order. Having consented to this representation of the system's dynamics as a function of two variables — differentiation (variety, complexity) and integration (order) — we still have to face the question of priority of these two variables.

Overall dynamics of the economic system — because it hinges on the scientific and technological progress — seems to point to divergence (differentiation), that is, growth of variety as the crux of the problem. Superficially, it might seem that the ideal set-up would assign priority to convergence and integration, characterized by equilibrium, with differentiation, which might lead to disequilibrium, relegated to a subordinate role. Simply put, it would seem advantageous if the scientific and technological discoveries could be elaborated and implemented according to a complete and consistent plan.

As a rule, economic models incorporating endogenous technological progress presume that it fits in with the overall picture in a complete and consistent manner. The cornerstone of these models is attaining the state of equilibrium, that is, integration; technological progress or differentiation is subservient to this task. In other words, the above scenario describes a situation where development has degenerated into growth.

Nonetheless, the nature of the creative process is such that it refuses to yield to a complete and consistent development plan. Therefore, the creation of new things revolves around the state of disequilibrium, and the task of the creators is to heighten disequilibrium. The opposite (to models with endogenous technological progress) kinds of models place the accent on technological progress, which introduces disequilibrium. Subsequent attempts to reach equilibrium, which alleviates disorder, are deemed secondary.

The economic ideas of Joseph Schumpeter surpassed classical economic theory by emphasizing economic disequilibrium, brought about by the random actions of entrepreneurs, and assigning the mechanism of coordination, designated to balance the economy and prevent its complete disintegration, a subsidiary function.[3]

Initially, structuring economic dynamics revolves around trying to attain equilibrium also with respect to the expansion of variety, that is, to reduce entropy. The cornerstone of the second approach is to promote disequilibrium, that is, growth of entropy or variety induced by spontaneous technological progress. Aspirations toward equilibrium are subordinate, designed to merely balance the system in order to keep it from exploding. As we can see, the second approach does not negate the importance of equilibrium. If technological progress were allowed to aggravate disequilibrium uncontrollably, the economic system might shatter. To prevent this it is vital to coordinate the activities of the participants, a task accomplished via all kinds of horizontal and vertical economic mechanisms. Generally speaking, the second approach is conducive to the growth of sytropy of the system as a whole, as corroborated by the experience of developed Western countries (Japan included).[4]

The various methods used to attain the best possible value of the sytropy function are discussed below.

BELIEF, AESTHETIC, AND SCIENTIFIC METHODS

Having expanded our arsenal of indeterministic ideas and having sketched the general form of the criterion used to measure the degree of indeterminism, we are now in a position to tackle the problem of finding the degree of indeterminism. Let me adopt an instrumental approach sometimes used by physicists in introducing the notion of measure. Changes in the structure of matter at a molecular level are associated with changes in temperature, that is, the speed at which the molecules move. Although there is a function to describe this relationship, physicists frequently choose to introduce various phases in the structure of matter (plasma, gas, liquid, solid states). From an operational perspective, each phase is distinguished by its respective phase-specific technologies.

I shall follow the physicists' footsteps and subdivide the spectrum of indeterminism into phases, each embodying its respective class of methods to solve problems peculiar to the given phase. These methods fall into three broad categories: belief, aesthetic, and scientific.

It may be supposed that these methods of cognition are located in the spectrum related to the measure of objectivity. Operationally, a method is objective if its results are reproducible, if it can be used by any qualified subject to generate similar effects from a similar set of initial conditions. Objectivity, being a matter of degree, can in principle be conceived as a continuum. But this continuum can have several phases.

I will point out three such phases from the viewpoint of difference in the methods of their realization. Two such phases are known — the belief and the scientific method. Correspondingly, in one extreme case the process is entirely subjective; in another it is completely objective.

The belief method encompasses situations where the network interlinking the constituent components of the system is incomplete and inconsistent. Indeed, within the framework of belief, it is impossible to discover the truth or falsity of perception. Matters of belief play an important role particularly in decision making in the area of scientific-technological innovations. Evaluations of such innovations are only partial because their side effects are unknown. Moreover, the more radical the innovation, the greater our ignorance of its implications and the greater the role of beliefs in its evaluation. For instance, the choice between an extrovert and an introvert course of development, that is, having technological progress or not, is in the end an act of belief, for neither can be proved superior to the other. Consequently, judging the consequences of certain actions (impact of one system upon other systems) depends solely on the subjective qualities of the operator. In summary one could say, regarding the subjectivity of the decisions, that the role of the belief factor gains with the innovativeness and radicalism of the idea. These situations are characteristic of complete indeterminism.

The scientific method works in those situations where the path between the system's present and future states can be fully and consistently described, at least probabilistically, both for the determining system and the determined systems. This task is accomplished by using objective criteria. The scientific method can be set up in an analytical form (in a form of a low) — as a production function, for instance, or in the form of an algorithm of the functioning of the system, balancing its input and output ingredients and the forces corresponding to them. But in all cases the probabilistic form, based on the discovery of frequencies, requires enormous information. The described procedure is characteristic of complete or almost complete determinism.[5]

The intricacies of both the belief and the scientific methods are well explored in the literature. Of interest here are the intermediary cases in the spectrum of objectivity and the special methods of action corresponding to them. At least one intermediary case is specifically associated with the aesthetic method. My discussion focuses on the aesthetic method.

The aesthetic method delineates situations where, on the one hand, different states of the system (including final states) cannot be glued together via the scientific method and, on the other hand, actions are not predicated solely upon one's beliefs. Under these circumstances, the impact upon development exerted by the system might be manifest in its beauty. This idea seems to be intuitively reasonable because beauty, without being a full and complete program for achieving desired ends, is still an ordered structure that creates predisposition to development.

In other words, under beauty the elements comprising the system being interrelated are based on something like the principle of nonpurposeful expediency — a principle rooted in an intricate blend of subjective and objective elements. The following quotation from Immanuel Kant (1966) illustrates my point: "Beauty is a form of worth of an object, which is appreciated without any idea of a goal" (p. 240). Georg Hegel (1968) elaborates on this idea: "Beauty consists in a visible realization of a unified corporeal appearance in its rest and motion and does not depend either on the object's utilitarian worth or on isolated accidental autonomous motion" (p. 134).

Thus, beauty is something that enables us, however incompletely and inconsistently, to determine the influence that the present state of development has on future progress. In this sense the beauty's progress represents striving toward the creation of completed (in the sense of maximum possible accomplished balance) incompleteness (in the sense of creating conditions for future development and disposal of dead ends).

Operationalizing the aesthetic method entails different procedures based on a reactive or selective search. Reactive methods emphasize the formulation of a set of rules (laws, heuristics) that, when used together, would ensure the realization of the potential; this approach resembles a search for the greatest harmony among the rules. The selective method tends to measure predisposition of an object toward development in terms of one parameter; this is akin to searching for something like the greatest beauty. The fact that selective and reactive methods flow into each other validates the statement that the aesthetic method deals with beauty.

My own approach to measuring indeterminism is based primarily upon selective methods rooted in the search for beauty. The working assumption governing this method is that beauty can be measured through its structure by first dissecting the totality and then gluing it back together. The described approach goes against the grain of popular opinion, which holds that beauty is holistic, so measuring beauty by carving up the whole is impossible because such atomization results in the loss of holistic vision. A mechanistic disassembling of the system's beauty into disjoint parts and subsequent reassembling of these parts would undoubtedly lead to the loss of the holistic effect. The proposed analytical representation of beauty is designed to combat this phenomenon by forming (during the partitioning of the entity) nontrivial structures that are amenable to subsequent integration with a minimal loss of the holistic effect.

The overall aesthetic evaluation must be to a great extent subjective because the evaluation itself is inseparable from the leader who is to realize this method. The above does not exclude the possibility of certain individual parameters being evaluated objectively.

It seems to me that our discussion concerning the measurement of predispositions endorses the feasibility of concocting such a procedure of analysis and subsequent synthesis of beauty. Everything that has been said regarding the aesthetic method is not meant to exclude either the scientific or the belief method but rather to limit the scope of their applicability forcing them to squeeze and make room for the third approach.

DIRECTING, POTENTIATING, AND PROGRAMMING

Constructive counterparts of the belief, the scientific, and the aesthetic methods are the methods of directing, programming, and potentiating, respectively.

Potentiating requires a special explanation. The aesthetic method is relevant to situations where at least the direction of the system's development has been selected but where the path from the present state of the system to some future desired one cannot be fully and consistently described. In this case creation of the system's potential is needed.

Potential is a measure of the system's predisposition to development. This approach is different from the probabilistic one that deals, albeit in terms of frequencies, with fully objective connections between the present and future states of the system. A potential forms a structure aimed at inducing the events in the environment in which the system is immersed to the system's advantage, preparing the system to channel unexpected outcomes in a way that is favorable to the system, and absorbing or reducing the shocks of unexpected events harmful to the system.

Such understanding of potential is also relevant for arriving at an evaluation of unique situations because it does not require knowledge of probability of use in the past of analogous situations. Hence the correspondence of the aesthetic method to the method of potentiating.

The directing, potentiating, and programming methods share some common features; first and foremost is the pluralistic mechanism. The more advanced case of the pluralistic mechanism typically presents a synthesis of the following stages: elaboration of different courses of action, selection of one (or a set) of them at each period of time, surveillance over the implementation of the chosen course, and replacement of the chosen course if it proves unsatisfactory. The pluralistic mechanism permeates all three methods of control (directing, potentiating, and programming). However, the implementation of each of these methods within a pluralistic framework raises some specific issues.

Directing is characterized by diversity of the proposed paths of development that are based upon one's beliefs. Because all the alternatives are a matter of belief, it is crucial to refrain from assigning priorities to the various alternatives. Under directing, the pluralistic mechanism is designed to preserve and expand the set of alternatives, that is, to support parallel existence of all the alternative courses of development, which originated from this directive method. Here, the various alternatives are not screened to determine the one that would govern the development of the system as a whole.

Religion serves as a good illustration of this type of design. Religion, rooted in the directing method, provides global vision, which governs the development of its adherents. One's religious views are rooted primarily in one's beliefs, although other methods, such as the aesthetic and even the scientific method (in certain branches of theology) come into play. In a free society, all religions coexist on an equal basis; religious institutions are separated from the state, and no single religion is chosen to govern the country's path of development.

Perhaps, different parts of the picture will come together after we examine the workings of a pluralistic mechanism under potentiating and programming. Under potentiating, pluralism manifests itself in the expanding set of alternatives of development; each is characterized by its own potential, evaluated by the proponents of this particular alternative. There exists an established procedure incorporated into the pluralistic mechanism, which governs the selection of a specific alternative (or a combination of different alternatives), picked to guide the development of the country. The implementation of the selected alternative is supervised by special institutions, and, if deemed improper, it is replaced by another program taken from an ever-expanding set of alternative programs. One crucial feature of this process is the constant interplay of convergent and divergent practices. In principle, it never comes to a termination point (never converges).

The pluralistic mechanism incorporated into the political systems of free countries presents an example of pluralistic potentiating. The existing as well as the newly formed political parties never cease developing new political programs. For the country as a whole to function during a given period of time, one alternative program (or their combination) must be selected; in other words, priorities must be established. The implementation of the selected program is supervised. In fact, it may be modified or (under certain conditions) even replaced altogether by another program taken from the set of programs being constantly elaborated by different political parties. Because of scientific-technological progress and a constantly changing environment, this process of program selection and replacement is essentially interminable, never attaining its final state when one and only one program remains.

The pluralistic mechanism can also be utilized under programming. It comes to the fore when there exist alternative methods of reaching the same goal and these methods are assigned respective probability estimates. Under these circumstances, it may be advisable to support parallel elaboration of different alternatives until one of them proves best (perhaps, in terms of probabilities), that is, until the process converges to a point. Having determined the best alternative, all effort can concentrate on the realization of this alternative until the goal is fulfilled. We observe this type of pluralistic mechanism. The solution of certain technological problems, when the feasibility of the various projects is already assessed in terms of probabilities, embraces this type of pluralistic mechanism. The so-called model of the two-armed bandit or, in its generalized form, the multiarmed bandit provides the mathematical model that explains the conditions under which parallel project development is preferable to one (Barry & Fristedt, 1985; Gittens, 1989).

Under complete certainty, the pluralistic mechanism disappears, and, from the very beginning, all effort is concentrated upon the implementation of the single best alternative.

The different forms of the pluralistic mechanism are not separated by a Chinese Wall. They transmute into each other under changing conditions. An alternative, initially rooted in a belief system, may begin to develop, eventually assuming the form of a potential and, subsequently, of a program. Accordingly, one mechanism of implementation gets replaced by other mechanisms.

The specific and the general features of the pluralistic mechanism were considered externally to the methods of directing, potentiating, and programming. Our subsequent discussion of these methods will be essentially one-sided, for we shall deal with the problems of developing and implementing a single alternative.

METHODS OF A SYSTEM'S OPERATION AND PLANNING

The objectives formulated above can be viewed within the scheme of the planning process (to which I have only introduced minor revisions) elaborated by Russell Ackoff (1981). Apart from its obvious application to economic planning, this scheme extends to any system (not just economics) as well as the various mechanisms of a system's operation (not just vertical ones that include planning). The conditions that give rise to different methods of planning might well result in different methods of operation of market-type mechanisms. This point comes to the fore if we consider the fact that anticipation takes place both in market-type systems and in planning. In the market, however, it is realized via the local interaction of the participants.[6] The actual practice of such preventive market-type tools as contracts and the operation of the stock market reveals the many conditions taken into account under planning. In fact, choosing the planning process as the basis for searching out the synthesis of various methods of performance of an economic system has another advantage, spurring a constructive search for the various methods of performance. The same investigation with respect to market mechanisms is more conducive to discovering the existing mechanisms.

A unique feature of Ackoff's approach to interactive planning has to do with his view of it as a multistage process. Planning methods at various stages are not the same; each stage is associated with different methodologies of planning. Moreover, the entirety of operational methods is not presented in a disjointed fashion: they are integrated into a single process. To this phenomenon we can give the name of integrative planning. It differs from uniform planning, which also represents a single process but employs the same methods at different stages.[7]

At the heart of Ackoff's approach to integrative planning lies the concept of purposeful systems that are capable of both generating their own goals and the means of reaching them through the course of their development. The concept of purpose-oriented is a key one here; it dominates all stages of the planning process.

This kind of purpose-oriented planning differs sharply from methods that place the heaviest emphasis on initial conditions and proceed by the extrapolation of past trends. A purposeful-oriented approach introduces into the planning process the idea of redesigning the future.[8]

The key issue in a multistage planning process is the formulation of the stages, that is, the specification of the conditions that induce shifts in the process requiring new methods of operation. Characterization of a given stage entails the formulation of its objectives, and, because the formulation is presented in terms of goal-directed processes, the immediate problem is the formulation of the criterion of development of a given stage. For the sake of brevity, goal and criterion of development will be treated as synonyms. Thus, the variety of stages manifests itself in the variety of the criteria of development. These criteria must fulfill certain stringent requirements, for they are called upon to ensure the unity of the process — to prevent local decisions from coming into conflict with the overall process of planning. It is highly undesirable to reach a goal that, for example, turns out to be an isolated, suboptimal result. The attainment of such goals leaves us with Pyrrhic victories.

Ackoff (1981) suggested that the planning process be divided into three stages, each assigned its respective criterion: ideals, objectives, and goals.[9] That is how he defines them:

> 1. *Goals*: those ends that we can expect to attain within the period covered by planning.
> 2. *Objectives*: those ends that we do not expect to attain within the period planned for but which we hope to attain later, and toward which we believe progress is possible within the period planned for.
> 3. *Ideals*: those ends that are believed to be unattainable but toward which we believe progress is possible during and after the period planned for (p. 63).

It seems to me other important categories termed aims and targets can be added to the above groups. These categories arise from the need to introduce intermediate goals directly interconnected with each other. The term "goal" will not do here because it is not directly linked with the corresponding objective.

Aims, in turn, differ from targets as far as the completeness of the set end. Whereas aims are specified in terms of all the important parameters pertaining to a given stage (for instance, profit), targets refer to individual parameters that are currently considered particularly valuable (for instance, output growth).

Aims and targets are distinguishable on all structural counts (for instance, by the degree of certainty and aggregation unmasked in the course of goal attainment). One area in which aims are well exposed is mathematical procedures of search for the optimum in multilevel processes such as dynamic programming. Characteristic of these procedures is setting goals for each individual step that would induce direct links between these steps. Aims are precisely this kind of step-by-step ends. Targets, however, are intrinsic to heuristic programming; in this case a planner is modifying aims on the basis of his intuition.

Constructing the described set of goals corresponds with different functions of the planning process and is accompanied by drastically different methods of operation.

Formation of ideals is related to the ends of the initial stage. The crucial function of this stage consists in setting a direction — directing — for an entire planning process.

Objectives of the second stage are less remote than ideals, and the means of reaching them are visible. However, this should not be at all construed to imply that from our position in the present we know precisely how to create a program for attaining such objectives. Therefore, it is important that we create initial conditions at the second stage that would facilitate the most efficacious development of the system in an unknown future. In other words, these initial conditions should be such as to make it possible to steer the future development of the system in a desirable direction, that is, create the greatest predisposition of the system toward future development in the direction of the chosen ideal. A structure that measures up to the requirements we set for the conditions in question was called a potential; and the process of its formation, potentiating.

Because objectives and goals, as they are defined by Ackoff, can assume the role of conditions for future activities unanalyzed in a plan, they should, in principle, be able to satisfy the requirements of our definition of potential. Here, the difference between the two notions consists of the fact that objectives themselves are potentials and that they are preceded by the potentials whereas goals are immediately preceded by aims.

Finally, the third stage is characterized by having goals that can contribute in a well-perceived manner to the realization of higher-level ends, as well as by offering a real possibility of generating programming methods for reaching such goals.

Possibly the term planning can be used to denote a certain metaconcept that fixes and synthesizes the above-mentioned methods of directing, potentiating, and programming. This kind of a metaconcept lays the foundation for subsequent attempts to relate the described methods of operation and their respective ends with the various modes of planning given from a contextual and temporal perspective. Thus, while using time-laden terms to describe various planning methods (strategic, mid-term, tactical, and operational) we ought to remember that we are concerned here with only a correlation between these methods and planning intervals. In a less turbulent environment, the methods of tactical planning may conceivably be extended to coincide with time periods equal to those that are currently associated with strategic planning.

Table 3.1 systemizes these interconnections. Two terms in the table should be clarified: the term "perspective" and the new complementary term "prospective." The difference between the two terms is

TABLE 3.1
Methods of Performance and Types of Planning

	Types of Planning		
Methods of Performance	*Context*	*Duration*	*Ends*
Directing	Prospective	Long-long-range	Ideals
Potentiating	Perspective Strategic	Long-range Mid-term	Objectives Goals
Programming	Tactical Operational	Short-term Current	Aims Targets

mirrored in their respective definitions. Webster's dictionary defines "prospective" as "looking toward the future" and "perspective" as "distant view." My other comment concerns the relationship between strategic and mid-term planning. The rest of the table is basically self-explanatory.

The finale to our investigation of the degree of indeterminism as it applies to system's operation (particularly, an economic system) is the analysis of the various styles of leadership.

TYPES OF STYLES — POSITIONAL STYLE

I chose to base the typology of various styles of decision making utilized by leaders on the methods of directing, potentiating, and programming. According to this classification, the various styles are distinguished primarily by the degree of objectivity of the leaders' decisions. This criterion gives rise to the following styles of decision makers, each with its own peculiar subvarieties: charismatic, positional, combinational, and connective.

A charismatic type of leader is matched with the directing method. Decisions derived through this method cannot be objectivized because the law of the excluded middle does not apply. The following definition of charisma (*Webster's New World Dictionary*, 1980) suggests the basic features of this style of decision making. "A special quality of leadership that captures the popular imagination and inspires allegiance and devotion."

Actions of a charismatic type of leader (problem formulation, its solution, as well as the consequences of his decisions) are typically evaluated based on faith in his impeccability. This style of leadership comes to the fore at a time of crisis when faith in the leadership is vital. The leader's personal charisma as judged and appreciated by his subordinates ensures absolute loyalty and even zeal in carrying out his decisions. Because much has been written about the charismatic style of leadership, I shall not dwell upon it.

Our discussion shall focus on the other two styles, the positional style and the combinational style. Because the two styles incorporate subjective and objective elements, they both correspond to the method of potentiating. The terms describing the two styles are borrowed from chess, so I shall use chess to illustrate them.[10]

Currently, chess lacks an optimal (mini-max) algorithm of play. It is impossible for players to link, in a complete and consistent manner, their current move with the final outcome of the game. As a result, the players have to proceed from start to finish without knowing for certainty the consequences of a given move. This uncertainty makes the formulation of intermediate goals pursued by a given move the crucial aspect of move selection. Many ingenious devices have been invented to deal with these problems.

The game of chess has been divided into several stages — the opening, the middle-game, and the end-game. These three stages outline the basic essence of these devices. For example, in the opening, the play is conducted based primarily on previous experience. In the end-game when the objective becomes clear because it coincides with the final objective of the game, methods of play might even incorporate such formal devices as optimization algorithms.

The greatest difficulties are posed by problem formulation in the middle-game. Currently, the prevailing style for the middle-game stage (for players of some caliber) is the positional style, which originated at the end of the nineteenth century and the beginning of the twentieth century. At the heart of this style is the evaluation of a position based on a set of material (essential) parameters and positional (relational) parameters, each weighed according to its respective unconditional valuation. This approach allows for a very elegant evaluation of a position because it allows material sacrifice for the sake of positional improvement (that is, positional sacrifice). In Chapter 8, I present a formal description of the structure of the weight function of a chess position and show it to be isomorphic to the representation of a position of a firm via the method of potentiating.

Taking the general case, the most characteristic feature of the positional style in chess is progression from the initial or given position toward the most remote positions within the limits imposed by the computational capability of the operator. The implementation of this procedure requires the operator to examine all the possible moves from a given position by conducting an exhaustive search of all possible variations. The first step of this procedure is to generate all one's own moves and all the legal replies by the opponent and then to evaluate all the generated positions. The first iteration of this procedure can be handled even by humans (although not under tournament conditions), not to speak of computers, because the number of different positions generated at the first step is about 1,000.

However, the position deemed best at the first iteration cannot be used to select the best course of play. The structure of the position, insofar as the set of positional (relational) parameters and their respective valuations, is insufficiently complete and consistent to determine the optimal course of the game. This means that the local optimum as deemed at the given step does not represent the global optimum; that is, the locally determined segment (move) of the trajectory does not belong to the extremum that characterizes the entire game. Consequently, raising the level of play entails the examination of all the possible moves as far in depth as possible. However, the enormous number of possible variations, approximately 1 million by the second move and a billion by the third move, makes such an exhaustive search prohibitive even for the most powerful and specialized chess computers, which probe only four to five moves deep.

As limited as the described search method is, once implemented in a computer (supplemented by a number of other methods and devices), it makes a formidable opponent for human players. Currently, the best computers perform at a level of a weak grandmaster (Hsu, 1990). A computer program superior to any player would probably require that the depth of such an exhaustive search be extended considerably.

There is a well-known dilemma between increasing the power of the computer, and thus extending the search horizon of all possible variations, and designing algorithms that, while not exact, would generate rough but effective solutions sufficient to achieve superiority of machine over man.

In the general case the evaluation of a position is based on a set of material and positional parameters all weighted according to their respective unconditional valuations. Generally speaking, a high degree of subjectivity is intrinsic to the positional style. It manifests itself in the subjective choice of relevant positional parameters and, especially, their evaluation. The objective component of positional style is largely derived from the evaluation of material parameters.

The subjective element arises not because the set of positional parameters and their valuations are formed based on a player's intuition. Rather, the choice of relevant parameters depends on the actual executor of the position, that is, the particular strengths and weaknesses of a given player. The role of the executor becomes vital because the actual realization of the position is not known beforehand, so future moves will have to be made based on the contingent situation at hand.

To sum up, the described method of evaluation of a position as the game's objective creates insurmountable computational difficulties. So, the way to improve chess algorithms is to modify the straightforward approach to positional style and attempt to generate satisfactory, if inexact, solutions by posing more limited goals and methods of their achievement.

One technique is to isolate a certain fragment of a given position, perhaps a rather vague fragment, but achieving that fragment is deemed sufficient to create preconditions for future victory. Here it is by no means apparent how the victory will actually be achieved. This limited-objective approach to positional chess is manifest in such narrow goals as the formation of a strong center or flank or in preparing an attack; these partial goals can be regarded as fragments of the overall position.

One important feature of positional style is an attempt to achieve the best possible value of the selected fragment represented as an aggregate parameter (in this sense, it is vague). This fragment of a position, which incorporates both material and positional parameters, shall be called a composite. By achieving the desired state of the composite we create predisposition for future success. One hidden assumption made here is that the overall structure of the position (other composites) is somehow taken into account; that is, the value of other composites does not decrease so much as nullify the gains in the value of the selected composite.[11]

Naturally, mere formulation of the overall goal in terms of a composite is inadequate. It is necessary to outline ways of achieving the posed goal. In chess, this broad outline assumes the form of a plan, basic guidelines governing play, which avoids exact calculation of the greater share of the variations leading to the desired goal, not to speak of exhaustive search. This style of play is associated with strategic vision of the game. This sort of strategy is predicated upon a sequence of intermediate positions that eventually lead to the desired position. Such aims as preventing the opponent's pieces from taking control of certain squares or putting and securing one's own pieces at the intended squares illustrate the above approach.

To sum up, a positional style (a style posing limited goals), incorporating composites as goals for a given stage, makes the search more economic as compared with the total search, which takes into account all the established parameters.

TYPES OF STYLES — COMBINATIONAL STYLE

Suppose that in analyzing a position, we are able to pinpoint certain primary parameters (material, for the most part) having such great value (in terms of unconditional valuations!) that achieving them would ensure long-run success. In fact, it is quite plausible (based on experience) to measure the isolated achievement of the intended magnitude of these parameters in terms of probabilities. The value of this isolated achievement might be very high. In other words, we can be confident that positions where such goals were fulfilled frequently lead to a successful outcome. Of course, it is indirectly assumed that the overall structure of the position is taken into account; that is, the value of other parameters does not diminish so much as outweigh the gained value of the selected parameter.

Elaboration of a program designed to achieve the desired magnitude of the selected primary parameter constitutes a special style of play that we call the combinational style. A characteristic feature of combinational style is the formulation of a narrow objective together with a completely specified program of its achievement (a rigorous chain linking the goal with all the intermediate stages). For instance, one typical goal posed by combinational play is to capture an opponent's piece and to elaborate a sequence of steps indicating a precise path toward that goal.

This style of play alleviates many problems by focusing on one narrowly defined goal rather than reconciling many different parameters, which is typical of positional play (even in forming such fragments as composites). Combinational style of play could be regarded as an extreme particular case of positional style because the former formulate the objective over a single parameter of the positional, and a material parameter at that. This aspect of combinational play is important for evaluating the material. As combinational style is described here, it is vitally important to realize that the evaluation of the goal parameter is unconditional, just like the evaluation of other parameters characterizing the position.

The subjective element of combinational style surfaces primarily in connection with target selection when all the parameters (except the selected one), both material and especially positional, are ignored. However, this approach can land one into trouble because positional parameters that possess some value are sacrificed in the name of material gains; that is, gains in material may result in an overall deterioration of one's position.

The actual elaboration of a program designed to achieve the posed combinational objective is an incredibly perplexing task. Because it involves all kinds of sacrifices and surprises, this style tests the player's creative potential, requires superior ingenuity, and is extremely difficult to formalize. However, once the program becomes known, it is largely amenable to external verification and can thus claim to be objectively correct.

Positional style neither repudiates nor abolishes the combinational one. The two are complementary: the positional style creates preconditions for future combinations. The combinational style presumes the conditions are basically ripe for a combination, so the task is to find the target and the means of attaining it. Positional

style, however, emphasizes the creation of preconditions necessary for a combination. It does not negate combinational style; it merely postpones it until the position is ripe for combinational play.

Another related aspect of the two styles of play is their aesthetic value. Both belong to the potentiating method, to the aesthetic method of a system's operation. According to popular opinion, positional style in chess is not as aesthetically appealing as the combinational style. (Naturally, this comment pertains to creative rather than routine play.)

Why does the combination style of chess induce such aesthetic intoxication? Perhaps, it abounds with peripeteia ("sudden and unexpected reversal of circumstances of situation in literary work . . .also in actual affairs" [*Webster's Third New International Dictionary*, 1986]), frequently involving sacrifices, which, in the end, prove to be justified. In comparison, the positional style of play, lacking violent sacrifices, seems dull and snail-like.

One well-known Soviet art critic (Volkenstein, 1931) who has pondered the aesthetics of chess has this to say:

> Beauty in chess springs forth from expediency of moves involving peripetia, from the unexpected and paradoxical (at least, at first glance) handling of a difficult situation. These are the moments when sacrifices take place. Victorious play conducted by means of gradual accumulation of tiny advantages without enchanting combinations could be called solid, well-composed, perhaps even instructive and fine, but there is no way it could be called beautiful (pp. 45–46).

However, let us not be so rash in judging the aesthetic merit of the various styles. To go beyond the superficial notions, we need to elaborate on the aesthetic method.

The criterion of beauty can be used to evaluate the state or result as well as the dynamics of the process leading to a certain state or result. The aesthetic evaluation of the state or result as it pertains to a position derives from the definition of beauty given above. This is precisely the reason why the evaluation of a position reflects the position's beauty: it reveals the position's predisposition toward development, and it captures in a nutshell the dynamics of the process, meaning the applicable parameters and their valuations express in a condensed form the capacity of the system to develop. The aesthetic value of the process is well described in the above quotation.

In fact, in some situations the result is evaluated instrumentally (pragmatically) while the process leading to it aesthetically is based on intrinsic values. Generalizing this situation, we see that there are four possible cases based on different combinations of instrumental and intrinsic values as they apply to the state or to the process. They are shown in Figure 3.1.

All four cases except 11 are offshoots of the aesthetic method. The most aesthetically complete case is 22 because aesthetic judgment governs the evaluation of both the result and the process. Cases 12 and 21 are aesthetically less complete. Of course, aesthetic completeness should not be equated with aesthetic impact impression.

In the light of the above statement, it becomes apparent that the positional style is aesthetically more complete than the combinational style. For one thing, the outcome of positional play belongs to the aesthetic realm whereas the result of combinational play is judged pragmatically. Therefore, it would probably be inconsistent to speak of the beauty of a combination, however wonderful it may be, if the resulting position fails to achieve the posed objective, which is largely amenable to instrumental verification.

As far as the process of goal attainment is concerned, both styles exhibit high aesthetic merit. Apart from discovering novel positional parameters of modifying the traditional valuation of the known ones, positional play includes what is called a positional sacrifice, which is very impressive as far as the element of surprise and its impact on the game. In a sense, positional sacrifices are more daring and risky than their combinational counterparts because a positional sacrifice of material is done for the sake of improving positional parameters; that is, the circumstances surrounding a positional sacrifice are such that neither the player nor an outside observer has a clear idea of how the sacrifice will be compensated. A combinational player proceeds differently: he thinks out the program beforehand to recoup the sacrificed material.

Because positional play strives for gradual improvement of one's positions by means of modifying isolated positional parameters (not to speak of ignoring the aesthetic merit of the outcome) and so produces minor peripeteia, popular opinion is that this style is unaesthetic.

FIGURE 3.1
Matrix of Combinations in the Aesthetic Method

Process	Results	
	Instrumental	Intrinsic
Instrumental	11	12
Intrinsic	21	22

Now let us leave chess and come back to the main topic of this chapter, the various styles of leadership.

One extreme case of combinational style, which in principle is amenable to complete objectivization shall be called connective. It corresponds to the programming method of operation. It is characteristic for this style that the formulation of both the goal and the process of goal attainment are in complete agreement with the general problem in which a given block is immersed. To accomplish this goal it is necessary to define a set of exclusively essential parameters in their totality. The evaluation of these parameters are current and situation-specific; that is, they absorb in condensed form all the diverse forces exerted by both essential parameters as well as the relations among them. A good illustration of this case is the classical approach to firm behavior. The first tenet of classical representation is that the firm strives for profit maximization as measured over all the input-output components for the period in which the discount function assumes values different from zero and that the firm can take current prices to achieve profit maximization, prices which bring in accord the firm's activities with its environment. Second, there is a highly detailed program for achieving the set goal, a program that can ultimately be represented as an optimization algorithm.

Our subsequent discussion will reveal that connective style contains two varieties. One entails supplementing the criterion of profit with certain qualifications upon some of its constituent components; the other does the reverse. It chops away from profit many of its constituent components and leaves only certain ones.

Let us in the beginning investigate the first case. One can assume there is an idealized, mathematically formalized mechanism governing the economy's operation that ultimately links completely and consistently the actions of all participants (firms, individual consumers, and the state). This mechanism is realized as an iterative process in which movement toward equilibrium proceeds as a linear approximation — via profit, expressed as a linear polynomial. Such a means of convergence, although monotonic, can take exceedingly long.

One way to speed up the process is to construct the profit criterion as a nonlinear function. However, constructing such a function is terrifically difficult. Therefore, the profit criterion has to be restricted to linear polynomial form, and all the possible constraints on output and input are added directly to it (*Decomposition of Large-Scale Problems*, 1973). These constraints are predominantly based on experience and thus allow system inertia to be taken into account. The indicators comprising these constraints are also those complementary to profit.

The other subclass of connective style is the combinational-connective style. In dealing with the general case, when introducing connective style, it was assumed that all the material parameters comprising profit (all expenditures upon input and the output) could be directly interlinked via prices. However, this is not always the case, and sometimes this approach is ill-advised. It may be impossible or inexpedient, for whatever reasons, to directly link all the individual parameters of profit with its overall value. Perhaps, the large number of variables makes the problem too cumbersome, suggesting we try to separate the set of variables which define profit and isolate those that can be determined independently. Subsequently, the arrived at result can be linked to profit. For instance, on the basis of market studies, a company that produces automobiles can determine the maximum feasible annual sales of cars for the upcoming years. This would enable the company to set the volume of production independently and possibly determine the structure of parameters that define expenditures by solving the corresponding optimization problem.

As the name suggests, combinational-connective style shares certain features with both the combinational style and the connective style. With regard to the combinational style, both formulate the objectives over a subset of essential parameters, such as output of a particular product or reducing some specific kind of expenditures (this corresponds to the term "target" introduced earlier). From the connective style, it borrows the type of valuations assigned to the components comprising the objective: the valuations (rather than being unconditional) are derived from the current conditions.

The category of connective style bridges the combinational style with both the positional and the connective styles. Within the framework of this scheme, I tried to deduce combinational style as one particular case of a positional style. However, combinational style gravitates toward the connective style, for it allows the various stages of the development trajectory to be interlinked in a highly complete and consistent manner. Moreover, because the combinational style treats material parameters as the important primary parameters determining the course of events, it becomes evermore apparent why the combinational style could be regarded as a borderline between the methods of potentiating and programming.

In charting the path of a system's development, it is important not to neglect positional parameters, which, often play a significant role — so significant as to allow for positional sacrifices (sacrifice of material in the name of improving positional parameters). Therefore, in imposing such narrow goals as gaining some isolated material parameter, we must always keep an eye on the value of positional parameters. This danger inherent in the combinational style should make clear why I chose to deduce it directly from the positional style.

To reiterate, the distinguishing feature of a positional style is its lack of a program establishing the value of intermediate stages so as to interlink these stages in a sufficiently complete and consistent fashion. Combinational style features a program that stipulates in a highly precise manner the very significant value of intermediate states over a narrow set of parameters (sometimes, a single primary parameter) and interlinks these states in a highly complete and consistent manner.

To sum up our discussion of leadership styles, it is reasonable to say that every leader of an organization combines, either consciously or unconsciously, all four styles described above. It would be desirable if every leader was proficient at all the styles, particularly the combinational style, because sooner or later he would have to resort to actions characteristic of this style. Of course, the above does not preclude a diversity of leaders each exhibiting his own dominant style of decision making.

In talking about the importance for leaders of an economic system (as a whole or its segments) to be acquainted with the various styles of management I would like to say a few words about teaching them these styles. It seems that the easiest style to teach, in terms of developing one's ability to construct primarily routine programs, is the connective style and its close companion the combinational style. This stems from the fact that these two styles are based on the programming (scientific) method and are more elaborated. It is not accidental that teaching business as business is rooted primarily in the scientific methods.

Much more intractable is the positional style and the goal-selection aspect of the combinational method. They are less routine and depend to a large extent on the subjective abilities of the leader. Instruction in this area must aim at enriching the leaders' intuition, helping them with conceptual recognition of what they are doing. In other words, learning the positional style could help leaders formulate and solve problems by enabling them to express explicitly the many intuitive considerations that were previously taken into account implicitly in the course of decision making.

Of course, the charismatic style of leadership is least teachable because it feeds on the subjective abilities of the leader. All these problems of teaching the leaders raise some sensitive social issues. Resorting once again to the analogy with chess, it is quite apparent that the achieved level of play is directly correlated to individual abilities. There is a limit to the level of play that can be achieved by teaching a person of average aptitude; beyond this level, teaching is practically ineffectual. It is well known that a positional sacrifice (not the routine kind) is not recommended for players below master's strength. But becoming a master takes more than just learning. Similar considerations apply in teaching the positional style (not to mention charismatic style) to business leaders.

My remarks are not intended to diminish the role of teaching in the business world; I only want to say that as the system of business education improves, we will continue to face some persistent problems (with all the ensuing social drawbacks) of matching the right kinds of leaders with particular kinds of tasks according to the leaders' individual abilities.

NOTES

1. A closed system possessing mechanisms of self-perfection can also strive toward greater negentropy. In other words, the growth of entropy is associated not only with the closeness of a system but also with the fixed nature of the rules governing interactions among its elements. Perfection is really a change in the rules of interaction. This may turn out to be equivalent to the openness of the system as far as the inflow of new energy is concerned.

My statement draws upon the idea that the total amount of potential energy of the system is determined on the basis of objects comprising the system and the mechanisms governing their interaction. The presence of the mechanism of perfection augments the total amount of energy in the system: while part of the total energy is spent supporting the operation of this mechanism, the system amasses a large amount of potential energy generated by the creation of new mechanisms of interaction.

2. This is precisely the way I interpret the following passage from the work of Lev Rozonoer (1973) on "resource dynamics," or the comparison between physical and economic models: "The first and the second parts of the work expound phenomenological models of equilibrium predicated on the existence of 'structural function' interpreted as entropy in thermodynamics and as the utility function in mathematical economics" (p. 116).

3. Similarly to Schumpeter's view of the role of entrepreneurs within the framework of horizontal market-type mechanisms, we can interpret the role of university scholars within the framework of horizontal nonmarket mechanisms.

4. In a private conversation, Israel Gelfand suggested that the approach he and a number of his associates elaborated — the so-called "ravined method" of solving multi-extremum problems — corresponds to the economic mechanism inherent in the second approach.

A reader interested in this analogy is referred to *The Quest* by Anatoly Pervozvansky (1970), which expounds the "ravined method."

5. Not by accident has Andrey Kolmogorov called the processes in which inputs are transformed in probabilistically ordered outputs stochastically deterministic. (See Dynkin, 1989.)

6. Quite possibly, the cardinal difference between a market and a bazaar is that the latter operates based on the principle of *post factum* (bring your product to the marketplace; if you sell it, good for you; if not, leave the marketplace). The market incorporates and responds to various mechanisms of anticipation.

7. The Soviet style of planning is a good illustration of what I have in mind. There, various planning stages are bound by temporal intervals (five-year, annual, quarterly, or monthly slices), but the planning methods for these intervals do not differ from each other in any significant way.

8. In 1974 R. Ackoff published a book bearing the same title, *Redesigning the Future*. It seems to me the proposed approach applies equally well to the reconstruction of the past.

9. As an aside, I would like to point out in this regard that the English language appears to be extremely well endowed with the vocabulary for expressing different shades of such categories. Thus, along with ideals, objectives, and goals, we also have purposes, aims, and targets. Possibly, all these categories can be brought together under the generic heading of ends. In my subsequent exposition, I intend to avail myself of the richness of the English language in order to introduce some new ends-related terms that may prove useful for describing certain aspects of the planning process.

10. The present discussion is an elaboration of Part 7, entitled "Chess as a Multistage Game," of my book, *Selected Topics in Indeterministic Systems* (1989).

11. Examples of such composites for a firm leader could be preparation for an assault on a competitor or strengthening one's influence in a certain region.

II

DIRECTING AND PROGRAMMING

4

Directing: On the Variety of Ideals and Ideologies

THE MEANING OF IDEALS

The introduction of the concept of an ideal into socioeconomic sciences has signified a very important development in theoretical thought — laden, however, with major drawbacks. On the face of it, the said concept resembles communist, fascist, and religious ideals that have been knocked off their pedestals by the wrecking ball of scientific methods. Noteworthy in this connection is the work of Karl Popper (1982). Popper has been very instrumental in showing that utopian engineering, or constructing a set of conditions sufficient to prove that these ideals are realizable, is implausible, so enlisting uncounted herds of people to implement these ideals is fraught with a bloodbath. Popper's alternative is to proceed in tiny individual steps (piecemeal engineering).

In my view, Ackoff's approach preserves ideals as beacons of development while dispelling all illusions that universal happiness will reign once these ideals are attained. The principal difference between ideal and utopia, however, is the fact that ideals can be changed and are inherently unfulfillable: we can come infinitely close to them without ever quite reaching them (Ackoff, 1981).

In this regard, I would like to suggest that great successes in science are not confined to providing solutions to problems posed — they also accrue to those who can show that a certain problem does not have a solution. From this standpoint, nineteenth-century science differs significantly from that of the eighteenth century. The 1700s witnessed a whole series of great scientific discoveries and inventions that produced in some a powerful illusion that purposeful and deliberate coordination of all these discoveries and inventions could offer boundless opportunities for bettering the lot of humanity.

The advent of the nineteenth century saw the appearance of monumental theoretical breakthroughs in thermodynamics that proved the impossibility of creating a perpetual motion machine. The hopes of many inventors to eradicate energy shortages were dashed. Instead of devoting their time and energy to finding that final stroke, which would make the creation of a perpetual motion machine a reality, scientists and inventors had to become meticulous sloggers striving for gradual increases in the mechanical efficiency of engines. The great discoveries in mathematics were to a considerable extent connected with impossibility theories. Thus, Evariste Galois proved that higher than fourth degree equations cannot have analytical solutions; Kurt Gödel proved the theorem of the impossibility of completing mathematics.

Along with natural sciences and mathematics, proofs of impossibility began to appear in the social sciences, for example, the theorem of Kenneth Arrow (1963) in which he proved that under certain circumstances there was no conceivable way for people to make general social welfare decisions in a way that could be considered democratic.

Still, social sciences in general are characterized by the persistent and erroneous faith that ideals constitute attainable ends: the existence of strong Marxist schools in the West seems to support the proposition that the category of impossibility has not yet been firmly implanted in the works of social scientists. Faith in human ability to create paradise on Earth by rational means is just as enduring as faith in paradise in heaven. The only difference is that the former possibility can be actually tested here on Earth. And even if experiments prove faulty, people will persist to believe in the possibility of heaven on Earth. Convinced they will not repeat all the mistakes of the past, people will seek new ways to bring this dream to reality.

In all fairness, I must add that claims of impossibility should, doubtlessly, be handled with considerable care: there have been some cases when even great scientists have mistakenly held to the infeasibility of doing something that later proved to be quite doable, for example, designing an airplane (Coffey, 1983). Moreover, in many instances, we can neither prove nor disprove whether something is accomplishable. In such cases, we have to accord equal weight to the rival claims of proponents and opponents of a certain course of action.

Still, my principal concern here is with a revolutionary role played by the concept of impossibility and an urgent need to see it further developed for application to social science.

What root causes make ideals unattainable? There are at least two reasons. According to the first one, an ideal state is unreachable simply because it runs counter to the fundamental principles on which our world is based. Under this category of explanation fall such things as perpetual motion machines and communism (if by communism we understand such a state in which all human needs will be satisfied).

The second reason refers to those situations in which a theoretical possibility that an ideal state may be attained is vitiated by the real-life patterns of action characteristic of a given system. This can be illustrated, for example, by the so-called retrograde positions in chess. In those positions, the actual location of pieces on the board does not violate any of the existing chess rules; but if we were to retrace our steps in order to see what sequence of moves might have led to one of those positions, we would discover that as long as we make only legal moves, none of those retrograde positions can ever be recreated.

It may be interesting from this perspective to look at the problem of disarmament. Theoretically, we can conceive a situation in which all weapons of mass destruction will be eliminated. However, the rather immutable human passions (judging by the literature, the same passions have endured from antiquity to the present) and the rules of interaction governing human relations have prevented the realization of this ideal since the beginning of recorded history. This does not mean that an ideal state of global and complete disarmament cannot be reached. It means that so far it is unclear whether the operational methods of interstate relations can bring it about and what new difficulties (for example, an all-round enslavement of people by some global overload) may follow the realization of such an ideal.

Being aware of the fact that ideals are unreachable has an enormous practical significance because it frees people from illusions that all they need is just one last push that would culminate in the solution of age-old problems and reward them for countless years of suffering. Those who use and misrepresent ideals as real ends realizable by means of practical action bring terrible misery to people supposedly exploited by the regime.

It is worthwhile in this regard to recall the dangers inherent in the ideals of communism. This ideal inspired numerous people in Russia, China, and other countries to carry out a revolution that they viewed as one last act of large-scale violence before the creation of society blessed with universal abundance. Marx taught that the leading industrial countries had already reached such a stage of development of their productive forces as to be immediately prepared to move into communism. As to underdeveloped countries, some Marxists thought that these countries could avail themselves of the advantages of socialism, improve their productive forces, and then usher in an era of communism.

On the basis of this interpretation of Marx, revolutionary leaders promised that the then present generation "would live under communism." The issuance of such promises was accompanied by grandiose plans for the transformation of their countries. In 1919 Lenin promised the Jung generation that they would live under communism. This pledge was backed by the state plan for the electrification of Russia. Karl Radec shrewdly called this plan "electrifiction." Wanting to create a military superpower in a short period of time, Soviet leaders declared that Lenin's pledge had to wait. In 1939 Stalin (1953), too, made similar pledges, but more cautiously. Soon after Stalin's death the leader of communist China Mao Tsetung officially announced that by the end of the 1970s the Chinese people would live under communism. With this end in mind, he launched the policy of the *Great Leaps Forward*. It can be safely stated that this policy halted the country's growth for about two decades. Heeding the call to glory issued by their younger brother (by that time, already claiming to be the older brother), the great Soviet brother represented by Nikita Khrushchev in 1961 announced his own program that charted out their path to communism. On behalf of the party, the first secretary solemnly promised that by 1980 "the present generation will live under communism." This kind of promise, undoubtedly, had to be backed up by some tangible figures. For the years 1961–1965, Khrushchev pledged absolutely unprecedented rates of growth (200 percent to 300 percent) of agricultural commodities, especially meat and dairy products, in the output of which the Soviets had at that time (and still continue to) lagged far behind the leading Western countries. But, since no serious measures were taken to ensure the promised rate of growth, the whole enterprise soon degenerated into all sorts of plots that took a very heavy toll on the state of Soviet agriculture. A temporary increase in the output of meat was achieved, among other things, by a wholesale slaughtering of cattle. We can easily appreciate the enormous damaged caused by this increase.

THE VARIETY OF IDEOLOGIES

The problem of variety of ideals is essentially tantamount to the problem of variety of ideologies. Ideology is here understood to be any coherent set of beliefs (axioms) regarding the direction of development of a given field. Properly speaking, the forces that determine the movement of people in a certain direction and by a certain distance provide the basis for a given ideology.

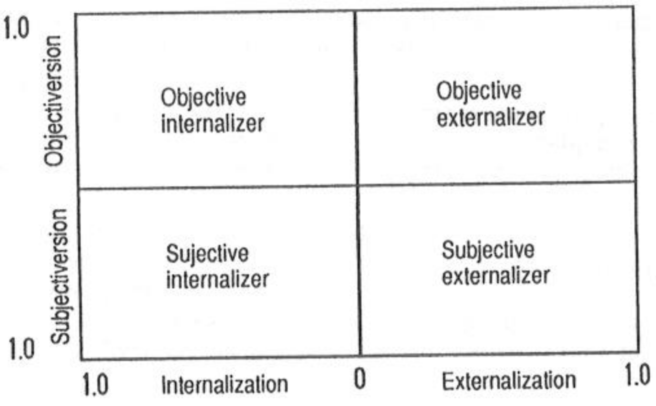
A preference for choosing this or that type of ideology can be regarded as an act of choosing a faith. Classification of personality types proposed by Russell Ackoff and Fred Emery (1980) greatly helped me to clear up this confusion. This classification is based upon an analysis of "the personality of an individual in terms of the two relations that can exist between him and his environment: (1) the extent to which his environment affects him (environmental responsiveness); and (2) the extent to which he affects his environment (environmental affectation)" (p. 117).

From the standpoint of an individual's response to environment, the authors distinguish between an objectivert (a person responsive to environment), and a subjectivert (one who is not responsive to environment). From the standpoint of an individual affecting the environment, the authors distinguish an externalizer (a person who is inclined to change the environment in accordance with needs) and an internalizer (one who, by changing the internal world, adapts to the environment). On the basis of classifying personalities in these two terms (that is, individual responding to the environment and individual affecting the environment), we can synthesize personality types. This typology is presented by a matrix (Figure 4.1).

The observed typology of personalities can be applied to a classification of different types of ideologies; we can name them analogously to personality types. The easiest types of ideologies for us to detect would be the extreme types. Thus, for example, Buddhism belongs to a subjective-internalistic group, and Marxism to an objective-externalistic one. Judaism apparently can be interpreted as a subjective-externalistic ideology, and Islam as objective-internalistic.

In this book, however, I have set for myself a narrower task of elucidating in a highly concise manner certain aspects of ideologies, which include the tendency toward changing the environment, called the extroverted as distinct from the introverted, which contain the tendency toward changing the internal world coupled with a rejection of change in the external world (Katsenelinboigen, 1984a, Chapter 3).

FIGURE 4.1
Matrix Representation of Personality Types



It seems to me that the history of civilization can be viewed, first of all, as an undulant changeover from extroverted ideologies to introverted ideologies and vice versa. Indeed, there appears to be sufficient evidence for this. It may be assumed that every region, in the course of its historical development, has experienced these periodic changeovers from one of the said ideologies to the other. Also, such changes continue to occur.

The Indian history, if we are to trust ancient books, stands out as the history of a country in which a changeover from an extroverted to an introverted ideology occurred in the times of Buddha and Asoka, and now the process of transition toward an extroverted ideology is again occurring in India. European history appears to be, first of all, a sample of the historical transition from the extroverted culture of Greece and Rome toward the predominantly introverted culture of the Middle Ages. The Renaissance in Europe is again connected with a return to extroverted ideology. In spite of the fact that this period also witnesses a return to some Greek trends oriented toward introvertedness (the works of Erasmus), this period is, nevertheless, dominated by an extraversive impulse. The latter was embodied in many forms, in particular in the moral foundation of bourgeois society as well as in Utopian designs.

The well-developed countries of Europe and America are once again facing the problem of transition toward an introverted ideology. A certain part of the young set in the developed Western countries, which has obtained a high standard of living and which has grown conscious of the futility of riches, seeks out ideals in introversiveness. The Hippie movement, which refused to accept the extroverted values of the older generation, was a symptom of this impulse. Early Christianity has become more popular, and interest in Buddhism is on the increase.

At present, the instances of scientists refusing to continue their work are still quite sporadic. Apparently, we cannot as yet talk of a definite tendency on the part of scientists to discontinue their research. But I would like to point out that the attention of scientists has been drawn to the baleful side of their activities in a rather blunt manner. Discontinuation of creative activities on the part of those scientists who produced new fundamental theories would pose an ultimate threat to ending technological development itself. If creative activities would cease, then the inertial development of existing ideas would be sooner or later bound to come to a grinding halt.

The reasons for such an undulant changeover of ideologies calls for an analysis of possible changes in psychology, as well as changes in environment. Let us remark that the more the people who lived in a given region were devoted to a concept of submission to nature, the more violent would be the character of a changeover to an extraverted ideology. Subsequently, the people, having made a transition toward an extraverted ideology, would discover the dangers that it bears. And then again the changes in ideology in the direction of introvertedness would occur.

I think that at this junction we can speak of at least four different dangers inherent in extraverted ideology:

- possibility of the earthly atmosphere becoming polluted due to massive waste of industrial production; depletion of the world's supply of natural resources and "depletion" of man;
- a threat of a holocaustal world war due to the creation of formerly unknown powerful means of mass annihilation;
- the study of certain new effects in connection with conducted experiments that might lead to the world's destruction due to the unavailability of appropriate technical means of controlling them (for example, new kinds of bacteria and chain reactions set off by atmospheric conflagration); and
- the creation of a new kind of species by biological (via genetic engineering) as well as nonorganic (via computers endowed with highly developed artificial intelligence) means in the course of evolution might threaten the existence of mankind due to the difficulty of inculcating into this species a set of necessary constraints.

Now let us see the dangers and the disadvantages of introvert ideology. At least, there come to mind:

- the threat of global disaster originating in outer space versus control over one's environment;
- people's fear of death and disease versus the desire for immortality in good health; and
- fear of starvation versus the desire to enjoy material well-being.

All these factors encourage people to strive for technological progress and, consequently, to an extravert culture.

Differences in the ideological development of various Western and Oriental countries exert a significant influence on their contemporary politics in economic development. On the basis of this viewpoint, we can now evaluate the ongoing debates in the developed Western countries, especially in the United States, concerning the admissibility of continuing with scientific progress. I first want to mention the debate in the United States concerning atomic energy and genetic engineering.

Eastern countries, owing to the upsurge of national revival and the danger of being invaded, have embarked upon an industrial development that entails a radical breakdown of the old introverted system of values and a subsequent transition toward an extroverted orientation. With that it is very important to bear in mind that the transition occurs differently in different countries of the East, depending on their prevalent traditions. In the case of China, for example, we can see an ongoing struggle (for at least 2,500 years) between extrovertedness (Confucianism) and introvertedness. It is no accident that in China there have been periods of various kinds of innovations, for example, the production of porcelain and the invention of gunpowder. In the case of Japan, we see a different kind of extrovertedness, which I would hazard to designate as cautious extrovertedness. Japan characteristically adopts innovations only after they have been tested in other countries. Quite a different situation exists in India, where introvertedness has been most thoroughly accepted. In India, industrial growth is rather slow, relative to the country's population. The desire on the part of Eastern countries to catch up on their industrial development within a short period of time often leads to the emergence in many of them of autocratic political regimes.

Had the developed Western countries been isolated, then the matter concerning the expediency of a further development of new technology could have been approached in a more affirmative way. The question becomes considerably more complicated when the aggressive countries, such as the USSR and especially China, undertake the development of new technology and then try to utilize its results for the purpose of conquering Western countries. Thus, we find ourselves on the horns of a dilemma: what poses a greater danger to the West — the potential danger of being destroyed by terantoid creatures, produced by genetic engineering, or by terantoid governments, created in aggressive countries? An answer to this question will depend (among other considerations) on the extent to which a general disillusionment with extroverted ideology will prevail in the West; how strong will be the West's genuine desire to combat these aggressive forces. Here we see a situation emerge quite reminiscent of the circumstances behind the historical decision made by the U.S. government to proceed with work on the atomic bomb in response to the threat that this bomb might be created by the German Nazis. It is hard to be a prophet in this world and to predict the outcome of the collision of these two waves in the contemporary development of the Orient and the Occident.

Thus, on the basis of what has been said, we can assume that the tidal wave of extroverted ideology, which has started to engulf the developed Western countries, coincides, as it were, with the tidal wave of introverted ideology engulfing the Eastern countries. It is precisely the collision of these two waves that makes such a curious picture of the modern world. Is not the Western world being corroded from within by its desire to stage a transition toward introverted ideology, subject to falling victim to the extroverted force roaring from the East? Was this not, indeed, the fate of Rome?

5

Programming I: Representing an Economic System

Let us imagine the economy as a system of consumption and production in which there exists a developed internal structure directed toward converting natural resources into final consumer goods. Changes occur in this system influenced by various factors and above all by endogenous technological progress. The economy is an open system that interacts with its external environment: the input of resources from its environment and output of resources into this environment may be observed in relation to its internal structure. Depending on context, we shall call individual goods and resources essentials or material parameters (objects). The entire process of transforming resources into goods we shall call input and output in real terms.

The structure of the economy, generally speaking, is such that it enables us to obtain a large number of different goods from a large number of different resources. The question of how the selection of the structure of external inputs and outputs will occur immediately arises in relation to this sort of system.

In the general case the construction of this selection mechanism requires an answer to the question of how the natural and human resources consumed by the system, as well as the products it outputs, will be valued, that is, the representation of a system by a pair of parameters — prime and conjugate. The value of a product we shall hereafter refer to as an attractor and that of a resource as a repulsor. Attractors and repulsors are both conjugate objects in relation to prime objects. The convertible resources and prepared goods, expressed in money terms, may thus be referred to as input and output in monetary terms.

The problem's formulation is linked to identifying value for inputs into the process and introducing methods for coordinating them. Because these values prove one way or another to be interconnected, only a part of them needs to be known; the other values may be unknown. With this in mind, the economy's functioning

represents a complex problem of balancing external known and unknown primary and conjugate objects with the system's internal configuration options. The different modes for representing the economy reflect the different approaches to achieving this balance under various conditions.[1]

THE FIRST HYPOSTASIS: PRODUCTION AND CONSUMPTION

Reactive Methods

If we assume that the economy interacts in a rather ordered and balanced fashion with its environment, then it may be represented as a system in which the links between initial objects are given, that is, between known resource expenditures and unknown levels of production in physical terms. Conjugate objects may be hidden away because it is assumed that the system will act like a black box in which these objects will be taken into account and balanced automatically. In other words, establishing direct connections between inputs and outputs does not eliminate the external forces that act on them but merely permits the system to be represented without their explicit expression.

These sorts of connections between input and output can be called reactive because they are reminiscent of a reaction in which a given input determines a given output. Reactive links, as we shall see, may be obtained by inductive and deductive methods. Let us start with the inductive approach, that is, from the assumption that they may be obtained on the basis of accumulating corresponding statistics. These statistics may be represented in the form of a corresponding tabular function.

If the resource expenditure element vector is delineated by the letter r and the consumer goods output vector by x, the tabular function will take on the following form:

$$r \to x \tag{I}$$

By virtue of the large number of possible variants of input-output, it is inconvenient to use a tabular function. Under these conditions the natural urge to find an analytical expression for the tabular function arises. Implementing this means finding a function that will match a given input with a given output. It is possible, of course, to construct a reverse function that will match a given output with a given input. However, for simplicity's sake, let us consider only an input-output function. We will call this function the production function and represent it as:

$$x = \pi(r) \tag{II}$$

The existence of a production function allows us not only to determine economically what predetermined amount of inputs has corresponded to what output, but also to project the possible output given the inputs. Thus, in contrast to the tabular function, the production function enables us to extrapolate.

Selective Methods

If the structure of the economic system remains fixed or sufficiently stable, then, all other things equal, a production function would be a reliable enough instrument for explaining the system's behavior. However, because of technological progress and all kinds of changes in the organization of the economic mechanism, on the one hand, and the changing representation of the values of inputs and outputs to the economy, on the other, assuming the structural stability of the system is inadequate.

Because of this effect, a new substance has to be introduced — system simulation, that is, the construction of an analogous system with which one may explain how the system will work when new technology appears. This analogous system must explain and select the quantity of goods to be produced and resources to be expended. Therefore the methods for solving the problems of finding unknown values for inputs and outputs can be called searching or selective. This kind of analogous mechanism assumes that the economic system be turned from a black box into a white one. In turn, the white box itself is based on black boxes that are, however, parts of other systems. These black boxes become, first and foremost, engineering structures. The links between their inputs and outputs are described by analogy with economics via engineering functions.

The distinction between engineering functions and production functions is as follows. Production functions reflect the internal organization of the given system, that is, depend on how people have organized the process, which is based on conditions external to them. An engineering function belongs to an object whose internal organization does not depend on humans because it is governed by natural mechanisms, which are external to the given system. Thus, for example, the physico-chemical processes that take place in a blast furnace will determine the links between the expenditure of iron ore, coke, and other resources to produce pig iron, slag, and gases. The function that describes these connections will be related to engineering functions. The term engineering function is a catchall. It includes as well agricultural functions expressing the link between, say, expenditure on fertilizer and crop yield.

Further, it is necessary to represent clearly the black boxes that make up the forces determining the inputs and outputs to the system. The environment's changing demands make this requirement particularly acute. It is a matter of clearly representing people's efforts and preferences, as well as the goods they produce and the resources taken from nature. This kind of clear delineation of attractors and their dependencies on corresponding volumes of consumer goods permits us to separate the more stable relations in the system from the more labile ones, which depend on technological progress. Based on this, the development of an economic system may be simulated, extrapolating only attractor changes. As for new technology and resources, their appearance will be taken into account on a situational basis.

Economics has two modes of representing this kind of simulated system. One of these could be called the differential approach, and the other, the integral. The two differ in how they describe the system's external forces. Primary objects and their linkages are given the same way in both approaches.

Differential Representation

Let us denote, as before, an unknown vector representing output of consumer goods by x. Let us now introduce forces that participate in the process as directly known values or, at least, as dependent unknown ones. Attractors will represent the marginal utilities for consumer goods. They may be directly known to us, but for generality's sake let us assume they are dependent variables: each marginal utilities vector will correspond to a particular bundle of consumer goods. Call this vector c. It is formed from the same external considerations as are engineering functions; that is, it is the result of sociopsychological processes occurring in people's minds. Let us denote the function connecting the marginal utility vector with the amount of consumer goods as follows:

$$c = f(x) \tag{III,1}$$

Function f in (III,1) is assumed known.

Let us denote, as before, a known vector representing the expended resources by r. Let us now introduce engineering functions as known technologies that connect inputs of resources with the output of production. For simplicity, let us linearize the entire range of engineering functions and represent them in the form of a predetermined matrix of input-output coefficients. The columns of this matrix show how many units of a given resource, using available technology, need to be expended to produce one unit of a given product. Call this matrix A.

In order to simulate the system, the law of conservation of matter must be adhered to. The requirement must be clearly expressed that

57

the amount of resources expended on production using available technologies may not exceed their supply. Formally, the law of conservation of matter in our notation will appear as:

$$Ax \leq r \qquad \text{(III,2)}$$

Let us now introduce restraining forces for the resources, that is, repulsors. They are unknown values. Let them be represented by the vector λ. Taking our new terms into account we may now write a law of conservation of energy for our simulated system as:

$$A^* \lambda \geq c \qquad \text{(III,3)}$$

where A^* represents a conjugate matrix of the input-output matrix A: the elements in its rows show how many units of different resources need to be expended to produce one unit of a particular product with a given technology. In this context the law of conservation of energy means that the value of resources expended cannot be less than that of the consumer goods produced.

Taking these relations into account, a differential model of the economy will be represented as follows:

$$\begin{array}{lll} c = f(x) & \text{(III,1)} & \\ Ax \leq r & \text{(III,2)} & \text{(III)} \\ A^*\lambda \geq c & \text{(III,3)} & \end{array}$$

It should be noted that although the number of relationships in (III) is equal to the number of variables, this system of inequalities does not fully determine x and λ. Here, it seems reasonable to place the following additional constraints on the vectors x and λ.

$$x_i > 0 \rightarrow A^*\lambda_j = c_i \qquad \text{(III,4)}$$
$$\lambda_j > 0 \rightarrow Ax_j = r_i \qquad \text{(IV,5)}$$

First condition means that if a good is produced then its marginal value equals its cost of production. We interpret the second condition in the following way: a resource having a positive value will be fully utilized. These two constraints in conjunction with (III) now form a system in which the number of variables is equal to the number of equations. Putting degenerate cases aside, the system we have thus defined is fully determined.

The differential model (III) is commonly referred to as the equilibrium model. I make use of the term equilibrium in order to clarify its definition from the point of view of the various different aspects characteristic to a systems approach.

From the functional viewpoint, the equilibrium model pursues the goal of finding values for unknowns in the system by which, according to the accepted development indicators, it may enjoy constant, unchanging growth (as a special case this might be a static state). Thus, under equilibrium not all parameters have

to remain identical. Far from it. Only those parameters that determine equilibrium remain unchanging.[2]

Thus, in the economic literature the term equilibrium coupled with the adjective dynamic is applied to models of economic growth. A parameter like the constant rate of growth is used to characterize the condition of equilibrium in them (Neumann, 1945).

In more complex economic models containing endogenous technological progress, dynamic equilibrium is characterized by the constancy of the controllable parameter that in the final analysis determines the distribution of resources between the expansion of productive capacity at current technology and the creation of new technologies (Zelikina, 1975).

From a structural viewpoint, the equilibrium model means introducing the panoply of primary and conjugate objects necessary to form laws of conservation of matter and energy. Only by simultaneously observing both these laws may equilibrium be described.[3]

From a process-oriented viewpoint, equilibrium implies mechanisms that will balance the primary and conjugate objects within the limits of the laws of conservation of matter and energy.[4]

Genetically speaking, from the standpoint of a view that engenders an approach to system development, equilibrium assumes the descriptive and causal representation of the system via a multitude of guiding counteracting forces.

To generalize, the economic system may be viewed as a kind of transformer of external inputs into external outputs. On the one hand, the transformation process takes place with the availability of all possible kinds of transformable and transforming materials, while, on the other hand, it takes place with the external forces acting as attractors to the system and those restraining its movement — repulsors. The value of the attractors (as well as repulsors) depends on the primary objects with which they are linked and may, generally speaking, even increase. In formal terms, this means the function linking the attractors' value and their primary objects is concave. For example, for a utility function this would correspond to the assumption that "eating makes one hungry." Pushkin's well-known fable about the golden fish also reminds us that the growing needs of certain types of people may also be described by a concave function.

The attractors' value might diminish as the amount of the primary objects with which they are linked increases. Formally, this is described by a convex function. Examples of this are the law of diminishing marginal utility for consumer goods and the law of diminishing factor productivity. In the latter case, the condition of saturation, when the maximum value of an attractor becomes equal to zero, may be singled out. Negative values for the first derivatives of this function are also possible, meaning that the system is beginning to lose ground because it has crossed the saturation threshold.

If the function is convex and has reached saturation, then development stops. Only the mechanisms maintaining the condition of saturation remain. Thus, equilibrium ought to be distinguished from saturation. Equilibrium is a special point. There the system has arrived at a position where the primary and conjugate objects are balanced within the framework of the laws of conservation of matter and energy. Saturation is also a special point. There the system has reached not only equilibrium, but also its limits of growth.

The gap between the real value of an attractor in the current state — call it relative — and its theoretically possible value — call it absolute — shall be termed development's driving force.

It could be that the existence of just such gaps in the value of attractors is the driving force behind the universe's evolution, including therein both the nonorganic as well as the organic world, human society, and self-evolving systems with artificial intelligence. The state of equilibrium merely implies that from the local perspective development is balanced with regard to both the primary and conjugate parameters. If the system has no opportunity for development, then the problem is reduced to just maintaining equilibrium.[5]

Integral Representations

Integral models in economics assume that one number is clearly being sought that will enable us to judge the effectiveness of the system's development. This number represents a value of the criterion of optimality. Let us denote the form of the criterion with F, its value with U, and its arguments as x. The system strives to optimize this criterion under the constraints set up by the previously mentioned law of conservation of matter $Ax \le r$. Let us call this kind of integral model an optimality model.

Thus, applied to the case at hand, the optimality model is expressed as follows:

$$\begin{array}{lll} U = F(x) - \max & (IV,1) & \\ Ax \le r & (IV,2) & (IV) \\ x \ge 0 & (IV,3) & \end{array}$$

After such a short characterization of the integral representation of a system, let us examine it in greater detail in light of the four aspects of a systems approach.

From the functional viewpoint, the integral, or optimization approach tends to express the level of a system's growth. The maximum or minimum that a system reaches also indicates that it has reached an optimal level of growth under the given conditions. Please note that the optimality approach allows us to measure clearly how much the system has progressed. If the optimality criterion's value is taken at two arbitrary points in time, then the change in its value may be used to judge the system's growth. With an equilibrium approach, for each of these points in time equilibrium states would be found and it would be impossible to directly compare which integral changes occurred in the system.

This kind of functional role in the integral approach can be implemented because, from the structural viewpoint, this approach directs attention to formation of an optimality criterion that integrates the driving forces. The presence of this criterion permits us in the final analysis to express the total value of the unknown variables in a single scalar expression. In equilibrium models the structure of the value parameters of input-output is given only in vectors.

As far as process is concerned, the integral approach assumes that the algorithms (or mechanisms) for seeking optimality and equilibrium can in principle be one and the same. The optimality approach offers wider girth for algorithm construction because fewer variables (in the given instance there are no resource prices) take part in the problem's direct formulation.

Finally, from the genetic standpoint, the integral approach implies a choice of position from which to view the problem that virtually assumes the existence of a leader or demiurge to institute a program of actions intended to realize his or her optimality criterion. By the same token, the integral approach is predisposed toward developing procedures for finding unknown quantities of outputs that will enable us to utilize scarce resources according to a predefined optimality criterion.

The integral systems concept, presented in model (IV), could be further developed. This more sophisticated representation involves two problems rather than one: prime and dual. Here, the ideas of optimization and game theory merge together. The concept of max-min surfaces in this connection. Under certain conditions, a saddle point may appear corresponding to the maximum in the prime problem and the minimum in the dual one. Therefore, let us denote previous problem (IV) as (IV)-1 and its dual problem as (IV)-2. The repulsors λ — resource prices vector — appear in this duality problem as unknown variables; all other known variables remain the same.

$$\begin{array}{lll} \lambda r - \text{min.} & (IV,1)\text{-}2 & \\ A^*\lambda \ge c & (IV,2)\text{-}2 & (IV)\text{-}2 \\ \lambda \ge 0 & (IV,3)\text{-}2 & \end{array}$$

As a whole, the initial and dual problems are represented in a linear approximation in the following form:

$$\begin{array}{llll} cx - \text{max.} & \text{(IV,1)–1} & \lambda r - \text{min.} & \text{(IV,1)–2} \\ Ax \leq r & \text{(IV,2)–1 (IV)–1} & A^{*}\lambda \geq c & \text{(IV,2)–2 (IV)-2 (V)} \\ x \geq 0 & \text{(IV,3)–1} & \lambda \geq 0 & \text{(IV,3)–2} \end{array}$$

The optimality criterion's value in the first problem, (IV,1)-1 is equal to its value in the dual problem (IV,1)-2 at the point of optimality.

Although the dual representation of optimality problems is a step forward, it is nonconstructive inasmuch as it is now unclear how to coordinate the pair of problems. This coordination may be directly accomplished via a simulation model or indirectly via an analytical representation. The Lagrange function is an example of this kind of analytical representation in which both primary and conjugate objects are simultaneously taken into account.

$$L = cx - Axp + rp - opt \qquad \text{(VI)}$$

This function has a quite clear economic meaning in the system of triple accounting. The expression cx reflects revenues the system receives from the sale of output, say, the value of consumer goods produced. Expenditures on the acquisition of resources, rp, make up the second bookkeeping entry. Cost of production, Axp, makes up the third. The requirement that cx, pr, and Axp equal out is nothing other than saying that, at equilibrium, the value of production must equal expenditures on the acquisition of resources and that the limited resources that are acquired are fully used up in this production by means of an optimal plan expressed in volume terms.[6] Correspondingly, on balance, the organization's earnings from the sale of output must equal both the cost of production as well as the expenditures on resource acquisition. It stands to reason that all these assertions make sense in a linear model.

Building a Lagrange function significantly eases the problem's solution, reducing it to classical methods. The complexity of solving problems of optimality management are essentially summed up in the impossibility of directly reducing them to a Lagrange function. This is because the problems' conditions are stated as inequalities: the Lagrange function is good only for solving problems of equalities.

THE SECOND HYPOSTASIS: SUPPLY AND DEMAND

Let us now examine an economic system in which the representations of supply and demand are clearly delineated from each other. Let all consumers be considered to act as one group, and producers likewise. Let us repeat all the same considerations in relation to them regarding the methods of representing the system that were earlier applied to production and consumption.

Initially, we may observe consumer demand for final goods — indicated by vector x ($x = x_1, x_2, \ldots, x_j, \ldots, x_n$) — and supply of these same goods by the producer — vector y ($y = y_1, y_2, \ldots, y_j, \ldots, y_n$). The parameter observed that determines the value of supply and demand for a product will be its price — indicated by a price vector, p ($p = p_1, p_2, \ldots, p_j, \ldots, p_n$).

To begin we are only able to construct a tabular function for supply and demand:

$$p \rightarrow x \qquad \text{(I,d)}$$
$$p \rightarrow y \qquad \text{(I,s)}$$

The problem is now how to equilibrate supply and demand. Having fixed a certain price we now try to find the values for supply and demand in the corresponding tabular functions that are equal.

Analytical supply and demand functions can be constructed on the basis of these tabular functions.

Let us represent the demand function as:

$$x = f_d(p) \qquad \text{(II,d)}$$

and the supply function as:

$$y = f_s(p) \qquad \text{(II,s)}$$

In this case, as is widely known, equilibrium between supply and demand will be achieved at the point of intersection of these two functions.

If we now open the black boxes that have determined the formation of demand and supply, each of them may be imagined as either an equilibrium model or an optimization model, relating to producers and consumers respectively.

$$\begin{array}{lll|lll} \multicolumn{3}{l}{\textit{Consumer Equilibrium Model}} & \multicolumn{3}{l}{\textit{Producer Equilibrium Model}} \\ px \leq w & \text{(III,d,1)} & \text{(III,d)} & Ay \leq r & \text{(III,s,1)} & \text{(III,s)} \\ p\omega \geq c & \text{(III,d,2)} & & A^{*}\lambda \geq \omega p & \text{(III,s,2)}^{7} & \\ & & x \leq y & \text{(III,d,s)} & & \end{array}$$

where:

p = vector of final goods prices
w = scalar value of consumer income

ω = scalar value of the marginal utility of consumer income
c = vector of marginal utilities of consumer goods for consumers
A = matrix of producer input-output coefficients
r = vector of producer resources
A^* = conjugate matrix of producer input-output coefficients
λ = vector of producer resource evaluations.

In case vector x has positive components, that is, $x_j > 0$ $p_j\omega = c_j$; if vector p has positive components, that is, $p_j > 0$ $x_j = y_j$; the last relations represent complementary slackness conditions.

In model (III,d) x (consumer demand for final goods), p (the price vector), and ω (marginal utility of consumer income) are unknown. In model (III,s), y (the output vector), and λ (the producer resource evaluations vector) are unknown. Because in the system of inequalities presented in models (III,d) and (III,s) the number of unknowns equals the number of inequalities, by solving the given system we can obtain balanced supply and demand. To be precise, additional constraints in the form of complementary slackness conditions must be introduced in order to achieve equilibrium. Further on, it will be implicitly assumed that these conditions are satisfied.

An optimality representation of producers and consumers is given as follows:

Consumer		*Producer*		
$u = f(x)$ – max.	(IV,d,1)	py – max.	(IV,s,1)	
	(IV,d)–1		(IV,s)–1	(IV,d,s)–1
$px \leq w$	(IV,d,2)	$Ay \leq r$	(IV,s,2)	

where:

$$w = py \tag{IV,d,3}$$

In particular, the construction of optimality model (IV,d)-1, describing consumer behavior, can be directly derived from an analytical representation of consumer behavior, that is, on the basis of the consumer demand function including in its arguments the vector of prices and consumer income. This assertion is grounded in the outstanding work of the economist Eugene Slutsky (1915), in which it is shown that under certain conditions, it is possible to turn the consumer demand function into an optimality model of consumer behavior.

$$x = f(p,w) \tag{IV,a}$$

Each of the above optimality problems, (IV,d)-1 and (IV,s)-1, may be formulated in accordance with a dual problem:

Consumer		*Producer*	
$u - f(x)$ – max.	ωw – min.	py – max.	λr – min.
(IV,d)–1	(IV,d)–2	(IV,s)–1	(IV,s)–2
			(IV,d,s)–1&2
$px \leq w$	$p\omega \geq c$	$Ay \leq r$	$A^*\lambda \geq \omega p$

LOCAL AND GLOBAL REPRESENTATIONS OF AN ECONOMIC SYSTEM

The initial model of production and consumption presented above assumed only a single, Robinson Crusoe-type participant. Its dual model counterpart could be interpreted in terms of game theory as a pair of players: one of them produces and consumes goods, whereas the other sells the resources he owns. Upon opening the black boxes, which had been concealed behind the figures of the buyer and seller in the supply and demand model, one could also assume the existence of two participants: producer and consumer. Although exchange between them had not been introduced, it was assumed that everything produced goes gratuitously to the consumer.

Meanwhile, in the general case, the economic system has a multitude of participants who enter into exchanges. This does not exclude the previous special situations, when there are participants who are only owners of resources or when all products flow free-of-charge from producer to consumer.

With relations of exchange the issue arises of the principles on which they are to be constructed. In essence, this is a matter of the principle to be used to synthesize the system. In the two hypostases of an economic system examined above, if more than one participant is allowed for, the following principle of system integration is operative. Consumption and production models are integrated like zero-sum games: what one consumer-participant (he is by the same token also a producer) wins, the other, the resource holder, loses. In the supply and demand model, essentially the same situation arises, but with a few differences. The consumer in this case only consumes whereas the producer, who possesses resources, must also produce, obtaining the maximum output he can. The output is simply taken away from the producer.

Thus, in both instances there are no exchange relations between the participants. I shall now radically modify the principle of system synthesis with several participants. Assume that each participant is simultaneously a consumer and a producer. Each of them has his own unique objective function and set of possible production alternatives; that is, he possesses particular technologies and resources.

Under these conditions, each participant may theoretically enhance his satisfaction if he enters into exchange relations with other participants. In order to protect some participants from the encroachments of other participants, it is assumed that not only all resources a participant has belong to him, but also the goods he produces and the income from them do so as well. From this stems the following principle of system synthesis, known as Pareto optimality: each participant may increase his level of utility so long as this does not decrease the utility of any other participant. Economic models with many participants can be constructed that are based on principles of system synthesis requiring income redistribution between participants. However, for simplicity I shall constrain myself to the Pareto optimality principle.

There are several ways to represent a system based on a Pareto optimality principle: local and global. By local I mean a system representation in which each individual participant is independent and the interrelations between them are based on certain rules. By global I mean a system representation where the emphasis is placed on the criterion of development for the system as a whole, subject to constraints on the feasible production spaces of individual participants.

Local representation of the participant set calls for an explicit insertion of local and global conjugate parameters, which denote an individual's values and the cementing forces respectively. An example of a local conjugate parameter is a subjective value of goods adopted by a given individual. The price of a product is an example of a global parameter.

From the systems perspective the price, as an integrator of an economic system, can be viewed in the following manner. Functionally, the price guides the development and allows an autonomous economic unit to choose the best course of action in determining its input/output when it is impossible for this unit to directly and operationally interface with its environment. From the structural point of view, the price of a given product reflects expenditures of different factors employed to produce it. From the standpoint of the process, prices represent an economic invariant; that is, it can be determined via vertical mechanisms (planning, for instance) as well as horizontal ones (such as a market). As far as genesis, prices evolved as an institution that helped the participants of the economic process avoid ineffective trade originating under barter-type exchange.

In the local approach, as in the models already examined, each individual participant, denoted by the index k ($k = 1, 2, \ldots, K$), can be represented in various ways; restrictions on participants are felt in that each of their incomes is based on the sale of what they produce. I shall discuss only two examples of this, equilibrium (model G1) and optimality (model G2), because they are important only to illustrate certain aspects of the problem of having a variety of modes of representing an economic system.

Local Equilibrium Representation

$$\begin{array}{lll} c^k = f^k(x^k) & \text{(G1,1)} & \\ A^k y^k \leq r^k & \text{(G1,2)} & \\ A^{k*}\lambda^k \geq c^k & \text{(G1,3)} & \text{(G1)} \\ p x^k \leq p y^k & \text{(G1,4)} & \\ p\omega^k \geq c^k & \text{(G1,5)} & \\ \sum x^k \leq \sum y^k & \text{(G1,6)} & \end{array}$$

Local Optimality Representation

$$\begin{array}{lll} u^k = F^k(x^k) - \max. & \text{(G2,1)} & \\ A^k y^k \leq r^k & \text{(G2,2)} & \text{(G2)} \\ p x^k \leq p y^k & \text{(G2,3)} & \\ \sum x^k \leq \sum y^k & \text{(G2,4)} & \end{array}$$

The equilibrium model (G1) is fully determined because the number of inequalities in it equals the number of unknowns. (Keep the complementary slackness conditions in mind.)

The global representation may be based on the internal or the external approach.[8] With the first one the principle of system synthesis introduced by the local representation is utilized. In this case, given sufficient conditions (first and foremost the fulfillment of relation (G3,3) as an equality, that is, the assumption that a participant fully uses his income), the local representation may be converted into the global one. Thus model (G2) may become model (G3):

$$\begin{array}{lll} \sum F^k(x^k)\mu^k - \max. & \text{(G3,1)} & \\ A^k y^k \leq r^k & \text{(G3,2)} & \text{(G3)} \\ p x^k \leq p y^k & \text{(G3,3)} & \\ \sum x^k \leq \sum y^k & \text{(G3,4)} & \end{array}$$

where μ^k (the weight of a participant in the system) is equal to $1/\omega^k$. Here, the μ^k coefficient denotes the inverse marginal utility of money generated by the sale of goods.[9]

The global criterion for the economic system may be dictated by external considerations. This criterion will express the demands of the suprasystem, which in the given example is human society and the natural environment.[10] Theoretically, a global criterion, developed from external considerations, is less arbitrary in the sense that it transfers the center of gravity of arbitrary assumptions, that is, acts of faith, onto the formation of prerequisites and the suprasystem's mechanism of performance.[11]

However, I feel it is important to stress not the origin of the global criterion, nor even the locality or the globality of how the economic system is represented, but the fact that in all representations of a system with many participants having a synthesizing principle is vital. Regardless of how this principle is given — in the form of rules of participant interaction (as it was in the local representation) or in that of global criteria — it will always have elements of arbitrariness. Which conditions and modes of system representation are more consistent with our goals is a subject for special investigation.

NOTES

1. The idea of researching the variety of means representing the economic system came to me many years ago while reading a book by Richard Feynman (1965), in which considerable attention is devoted to examining the variety of means of representing the physical system.

2. This is how equilibrium is defined by D. Moffat (1976):

> A stable economic condition in which there are no forces to prevent continuation of the condition. It is important to note that equilibrium is not the same as static; growth of an economic series is in equilibrium if it is changing according to its long-run pattern and can continue to do so, whereas *static* implies no change at all.

3. Equilibrium is sometimes defined only in relation to either primary objects or conjugate ones. Thus, for example, *Webster's Dictionary* (1980) defines equilibrium as either "a state of balance or equality between opposing forces" or "the condition in a reversible chemical reaction in which the products of the forward or direct reaction are consumed by the reverse reaction at the same rate as they are formed and there is no net change in the concentrations of the products or the reactants."

4. It is precisely this approach to defining equilibrium that is taken from the book *Time, Uncertainty, and Disequilibrium* (1979): "An equilibrium situation is one that has no tendency to change and that can be derived logically from a model that incorporates the operation of opposing forces. Stationary, dynamic, stochastic, and fluctuating (and so forth) equilibria are merely special cases of the more general concept" (p. 4). Webster's dictionary proposes an interesting procedural definition of equilibrium applied to certain physical processes: "the stage of a radioactive material at which the rate of disintegration and the rate of formation are equal."

5. In biology the conflict between proponents of homeostasis and the vitalists exactly reflects the correspondingly opposing points of view regarding maintaining a state of equilibrium and that of the sources of biological development.

6. This parallels the law of the conservation of energy: potential energy transforms into kinetic energy, which, in turn, transforms into potential energy in its new state. If kinetic energy were to vanish, the conservation of energy law would be invalidated.

7. The need to correct relation (III,s,2) for coefficient ω emerges from the following considerations. In the final analysis, the results (the values of the unknowns x and y) in the production-consumption and supply-demand models must be the same with $p_j > 0$. In the production-consumption model $A^*\lambda \geq c$ (III,3), but in the supply-demand model $p\omega \geq c$ (III,d,2). At equilibrium, for all $p_j > 0$ these inequalities become equalities. In order for these relations to be satisfied, $A^*\lambda$ has to equal $p\omega$.

8. These considerations regarding local/global and internal/external parameters are also pertinent in the analysis of the sociopsychological frame of an individual and his interaction with other individuals. Perhaps man is regarded as a microcosm or as an individual capable of autonomous decision making because his emotional mechanism, his physiological field, incorporates not only local values that reflect his selfish interests but also global values that embody his unity with the outside world. It seems these global values are precisely what we customarily call conscience. (Yury Koriakin has called my attention to the etymology of the word "conscience." It consists of two parts, "con" and "science," and seems to reflect knowledge of one's link to the outside world. The word "conscience" is a calque of the Greek word. The word traveled from Greek into Latin and subsequently into German. It found its way into Russian with the translation of the Bible). The totality of these local and global values forms an individual, that is, a human being who combines his selfish interests with the interests of others. Naturally, internal local and global values are accompanied by their external counterparts. For instance, moral statutes that extend to all people and the laws represent this kind of external global parameter. The individual's decision making is based upon all these internal and external values. As the person makes numerous decisions, he evolves and accumulates experience that manifests in his judgment of various outcomes or actions. But these values belong to the personal psychological realm.

9. Primary parameters given in physical terms and the corresponding conjugate parameters are assigned zero rank. Parameters with rank one arise when the value representation is achieved as a result of direct transformation of the physical magnitude via the conjugate parameters. The conjugate parameter corresponding to this value representation also occupies rank one. py^k and μ^k are examples of primary and conjugate parameters of rank one, correspondingly.

10. Such an approach to forming a global criterion somehow stands against the internal approach that assumes the principle of system synthesis is conventional or based on the power of one group (or, at the extreme, a single leader) over all society.

If the variety of people's opinions are taken into account, then working out a system synthesis principle by conventional means is quite difficult. Here either voting mechanisms or force comes into play. There are always elements of arbitrariness.

11. If we continue the chain of suprasystem construction for the economic system, then it will be necessary to build a suprasystem for human society. This may be the universe. One could suppose that the system "human society" occupies an intermediate spot between the organic world and the new artificial world being created in which artificial intelligence will rule.

Such an approach to the external criterion for human society does not exclude consideration of the particular emotional make-up of individual persons. These personal values become limiting factors; more precisely, the global criterion earlier alluded to is based on the synthesis of these individual objective functions. Inasmuch as the external and internal criteria, generally speaking, may not correspond, stimulatory operators must be developed to place the internal global criterion in correspondence with the demands of the external global criterion. It is a special case when the external and internal global criteria may coincide in the sense that they both define the same path of change of human society.

6

Programming II: Interpreting the Variety of Modes for Economic System Representation

The different modes of economic system representation we have looked at do not stand in opposition to each other but are instead complementary. The possibility of converting one representational method into another and their support for each other uncover new ways to form more effective mechanisms of economic performance. Difficulties arise in explaining the conditions for applying this or that representation of an economic system and their synthesis. Later on I shall illustrate this with examples of combining different representation modes by diminishing their commonality.

COMBINING THE TWO HYPOSTASES OF ECONOMIC SYSTEM REPRESENTATION: MULTIPLE PRODUCERS AND CONSUMERS

In models (G1)–(G3) the right-hand parts of the relations (G1,4), (G2,3), (G3,3), denoted by py^k, actually represent the income of participant k. In models (III,d) and (IV,d)-1, consumer income denoted by w belongs to the right-hand part of the relations (III,d,1; IV,d,2).

The category of income fulfills a vital function. It presents a consumer with an opportunity to act autonomously, that is, to pick the desired basket of goods priced respectively under such global constraint as the available income. If prices and income are lacking, consumers must be given all the produced goods directly. This situation is characteristic of extreme scarcity of resources, for instance, in concentration camps or at the time of a war when distribution is conducted by means of a rationing system. Growth of output of consumer goods and the resulting rise of the standard of living lead to a flourishing shadow economy in case the rationing system is not abolished. It reallocates the acquired goods according to the preferences of the participants. This mode of distribution is wasteful, so another system based on prices and incomes is introduced (Katsenelinboigen, 1978).

The assumption permeating the said models was that the income of each (collective) participant (prices are assumed known) would lead to picking a set of goods that corresponds to the state of equilibrium (optimality) for the system as a whole. As a preview, I want to note that in the real world the economy may not be close to the state of equilibrium. Other aggravating unknowns are the introduction of new products on the market and the uncertainty of the volume of production of the existing goods and their respective prices. But even in this case, income represents a universal constraint that delimits the number of choices in the basket of goods preferred by a given participant.[1] The economic system is ordered as income is adjusted over the course of the system's operation or as shifts in output take place or as prices change.

Combining Two Hypostases of Economic System Representation: An Averaged Consumer

For simplicity at the start, the representation may be limited to the consumption-production hypostasis. Such an approach to system examination is particularly useful in explaining the possible distributions of resources when technological innovations appear. In implementing such an approach, a global optimality representation of the economic system is especially valuable. But in order to give shape to this, it is necessary to clarify a global optimality criterion for development. How can we obtain this kind of criterion?

Theoretically it is possible to assume that this criterion must reflect consumer preferences. It is thoroughly possible to further assume in the first approximation that all consumers are represented as one statistical aggregate. This assumption, in turn, is based on an act of faith that, statistically, consumer preferences are largely sensible, if only because they reflect the development of evaluations. People's feelings, apparently, have been formed in the course of evolution in such a way that their evaluations of different objects and actions run in the direction of an evolutionary process.

Thus, if we assume in the first approximation that the economy's optimality criterion is oriented toward growing prosperity and preclude a criterion that radically breaks with established tradition in human preferences without any basis for doing so, we can construct the criterion as follows.

Let us examine the market mechanism with the assumption that, statistically, consumers, as an entity, behave *in toto* rationally: in accordance with established prices they allocate their incomes so as to maximally satisfy their needs. On the basis of accumulated statistics, consumer behavior can be constructed in a tabular function in which certain prices and income will be set in accord with a certain level of demand for consumer goods. On the basis of this tabular function it is possible to construct a demand function. In turn, as follows from the work of E. Slutsky mentioned earlier and subsequent research stemming from it, an equilibrium model for consumer behavior can be built based on our knowledge of the demand function. Under the right conditions this model can be converted to an optimality one with a manifestly expressed optimality criterion. This criterion is global by virtue of the assumption made above about representing consumers as a unified statistical aggregate.

The preceding permits us to assert that adding together two hypostases of system representation, consumption-production and supply-demand, gives us the chance to expand the framework of investigating the mechanisms of economic system performance.

Combining Reactive and Selective Representations

Both the reactive and selective methods of system representation similarly do not have to be placed in opposition to each other but coexist in complementarity. Each of these methods has its advantages and disadvantages and may be used with this or that system element depending on conditions.[2]

Thus tabular function methods and analytical presentations enable us to find problem solutions more quickly than by methods involving balancing the system with equilibrium-optimality concepts. Analytical concepts open new opportunities for qualitative analysis. In turn, the equilibrium (or optimality) concept enables us to take better account of innovation.

Thus the combined use of these methods allows us to furnish the best possible system performance. Of course, synthesizing the various different methods of system representation is not so simple. It is difficult to determine the boundaries of their useful application. These difficulties stem from the fact that it is not clear how to measure the cost-benefit of each of these modes. The most suitable method of measuring cost-benefit dictates the best mode of representation of a system. From this it follows how important it is not to unify any one of these methods, regardless of the attractiveness of such an operation: each of them offers particular advantages. It follows as well how important it is not to oppose these methods to each other but to find ways to synthesize them.

The above interpretation of the kinship between reactive and selective methods (taking into account the proposed definition of indeterminism as the degree of incompleteness and inconsistency of the process) poses a challenge to the widespread notion, also expounded by Russell Ackoff (1981), that teleological description implies "an output-oriented way, rather than deterministically, in an input-oriented way" (p. 21). In other words, an indeterministic vision of the economy is invariant with respect to the mode of the system's representation. Both methods can be deterministic as well as indeterministic, which all depends on the degree of completeness and consistency of the methods' constituent links rather than on what actually defines the process — input or output.

Combining Equilibrium and Optimality Concepts

Comparing systems representations using equilibrium and optimality models presents difficulties. In using an equilibrium approach, we encounter colossal problems. They have to do with elaborating axioms and concern the inequality/equalities and variables sufficient to obtain a fully defined system. Even in the case of model (III), with only one participant, it is quite hard to define condition (III,3) as an axiomatic requirement. It is by no means apparent that this condition represents a requisite addition to the self-evident inequalities governing the allocation of resources (III,2). In the optimality approach the axiom may be obtained as a theorem pertaining to the necessary conditions of optimality.

In the more complicated economic models, where there are many participants, elaborating this kind of requirement is much more difficult. As follows from the previously introduced equilibrium model of an economic system with many participants entering into exchange relations, new variables and inequalities, which are quite hard to formulate on the axiomatic level, have to be considered. I have in mind particularly relation (G1,5) and the variables ω^k included therein. Meanwhile, from the model of consumer behavior represented in optimization form in (IV,d)-1, this relation can be obtained on the theorem level as the necessary condition of optimality.

Equilibrium models, in turn, naturally assume axioms that, it can be shown, are transformable under certain conditions into the relations used in the optimality approach. Thus, it can be shown that under certain conditions the differential form of (III,1), which describes the marginal utilities of consumer goods in relation to their volume, may be transformed into an integral form represented by (IV,1)-1, which describes the optimality criterion.

This transformation of equilibrium models into optimization ones can be quite useful in analyzing equilibrium models, inasmuch as the optimization approach has its advantages. Thus the existence of equilibrium in Walras's model (model III) was demonstrated in the 1930s by Minkovsky without using the idea of optimality. But this proof was so complex and cumbersome that even professional mathematicians had difficulty understanding it. A comparatively simple and elegant proof of the same problem was given after World War II. This became possible following the development of a mathematical programming device and, in particular, the use of the Kuhn-Tucker theorem connected with this.

THE LOGICAL AND HISTORICAL PATH OF MODES OF REPRESENTATION OF AN ECONOMIC SYSTEM

It is apparent from the different ways of representing an economic system looked at earlier that these representations could be obtained by an inductive as well as by a deductive approach. The inductive path is expressed in the following logical sequence:

Based on the tabular function (model I), an analytical representation in terms of primary objects may be obtained without explicitly introducing conjugate objects (model II).

This representation renders a proximate solution of the problem because both initial information and our ability to approximate the existing statistical data by appropriate correlations are limited.

This analytical representation, in turn, may, under the right circumstances, enable us to obtain an equilibrium representation (model III).

If the latter's differential form is integrable, describing in this case the link between marginal utilities and the amount of the goods consumed (III,2), the equilibrium representation may be transformed into an optimality one (model IV).

This optimality representation may be expressed by a pair of max-min problems (model V:IV-1 and IV-2).

Under the right conditions, the latter may be expressed analytically in a Lagrange function (model VI).

By the same token, under certain conditions, it is possible to reverse engineer a deductive sequence of methods for representing the economic system.

Knowledge of a higher level of representation permits us to represent the next level. Thus from the analytical Lagrange function (model VI), in which primary and conjugate objects are simultaneously represented, a pair of optimization max-min models (V) may be obtained. From these optimality models (IV), in turn, we can arrive at an equilibrium model (III). Analysis of the equilibrium model enables us, in certain cases, to obtain a general solution, an analytical representation of the system in terms of its primary objects (model II). This representation possesses certain advantages over the equilibrium one. It makes possible a qualitative analysis of the corresponding problems because it permits us to see very clearly the connection between system inputs and outputs. At the same time it significantly eases the task of quantitative analysis, that is, obtaining the proper quantitative values for unknown variables. Knowing the analytical function, a tabular function can be easily constructed and used as a convenient means in those instances when it would ease calculation. For example, the multiplication table and the logarithmic table exist because of analytical procedures corresponding to these calculations.

The analytical representation we have looked at, because it was obtained deductively, has a special value compared to the analytical representation derived inductively. Precisely when these two types of analytical expressions are in close enough correspondence, we can be more certain that the law found by empirical means (inductively) is sufficiently well-grounded because it has theoretical (deductive) corroboration.

In other words, the preceding argument can be interpreted as follows. Any deductive theory, before it is applied to existing conditions, must be experimentally verified. Creating an adequate experimental basis, that is, not having to test major projects on live human beings, is one of the most complex and current problems in the social sciences. As for inductive methods, they may be experiential as well as experimental. If they are based solely on past experience, they are not the outcome of controlled conceptual verification; they are merely experiential. These laws are quite important and may be especially useful in the social context for comparing the expected results of new projects with past tendencies. Characteristic of the experimental inductive method is the experiment's orientation toward verifying the stated hypothesis.

In all cases where a solution exists in the equilibrium model, it can be solved numerically, which would correspond to the direct formation of a tabular function of primary objects (model I) without an earlier general solution of the problem, that is, construction of an analytical expression.

In sciences enjoying a long history, and above all in physics, the appearance of new modes of system representation has taken place first and foremost inductively. Deduction has accompanied induction but primarily in the reverse direction, that is, from new system representations, inductively arrived at, the old ones were deductively reconstructed. Thus the development of physics, starting with the Renaissance, has primarily been expressed in the derivation of analytical representations, that is, empirical laws, from tabular functions constructed on the basis of experimental data.[3]

Johan Kepler (1571–1630), on the basis of tabular functions compiled from the results of many years of exact astronomical observations by Tycho Brahe (1546–1601), inductively arrived at three laws for planetary motion (for example, the first law states that a planet's orbit is elliptical). All these laws completely neglect the forces behind planetary motion.

Isaac Newton (1642–1727), explicitly introducing forces, proposed an active approach to researching the processes of motion. His equilibrium model was based on fundamental principles of the universe's development or primary laws. (These principles are known as Newton's three laws. One of these, for example, states that every action has an equal and opposite reaction.) Inasmuch as physics is the fundamental science about the universe, laws (or principles) pertaining to the universe are associated with it. Therefore physics' initial laws are like primary ones, rather than laws obtained from a suprasystem. From Newtonian equilibrium models, it has been possible to deductively arrive at laws of behavior for individual physical systems. Newton himself did this, deriving from his general model the aforementioned three laws of planetary motion discovered by Kepler. Thus the primary laws of physics were built into the foundation of Newton's model. It was on this basis that analytical representations of the behavior of individual physical systems, formerly based on empirical data, were derived already as secondary laws of physics.

Subsequently, an alternative principle of mechanics, proposed by Pierre Maupertuis (1698–1759), enabled us to find a new manifestation of the active approach to representing physical systems. Leonard Euler (1707–1783) managed to demonstrate that the extremal principle could, under proper conditions, lead deductively to an equilibrium model. Joseph Lagrange (1736–1813) subsequently proposed a new mode for representing the alternative problems with a special function he developed that used special multipliers; this function, like its multipliers, now bears the author's name.

By this discussion of the sequence of the development of methods for deriving laws in physics I by no means intend to imply that in physics, thanks to progress in the construction of equilibrium and optimality models, the need to find secondary laws directly on the basis of experimental data has vanished. The relative simplicity of secondary laws, their transparency for qualitative analysis, and the ease with which they may be quickly calculated have earned them well-deserved popularity not only among physicists but also among all natural scientists. However, while broadly constructing these kinds of laws, their authors understand that they are not based on absolute truth. These laws are more constructive the greater the number of cases to which they may be applied or the larger the scope of the general theories to which they correspond.

Economics is one of the younger branches of science. For this reason exact modes of system representation, analogous to those I mentioned in physics, appeared relatively late, at the end of the nineteenth and beginning of the twentieth centuries. The appearance of new methods took place rather sporadically, possibly partly because of the influence of other already mature sciences, above all the physical ones with their arsenal of precise modes of system representation. Thus, Leon Walras's equilibrium models preceded analytical representations, expressed as production functions. The optimality representation of an economic system in local form, known as the Pareto optimality model, came in a period between equilibrium models and analytical representations. At approximately the same time, Enrico Barone, Eugene Slutsky, and Frank Ramsey developed a global optimization model. Finally, the 1940s saw the introduction of the most developed dual representation of the economy through game models by John von Neumann and Oscar Morgenstern. Constructive procedures for solving optimization problems furthering this mode of representation were elaborated in the work of Leonid Kantorovich, Tjalling Koopmans, and George Dantzig.

All the equilibrium and optimality models mentioned so far were based on primary laws taken from a suprasystem — society. These laws are those of human psychology, on the one hand, and, on the other, production possibility frontier limited by the discovery of natural resources and engineering functions.

If the appearance of new modes of representing economics had not progressed as sequentially as in physics, their mathematical analysis and practical implementation (numerical reduction) did repeat physics' path. Thus, the Cobb-Douglas production functions, based on empirical material, were constructed in the 1920s using econometric methods. The derivation of other economic laws using the same econometric methods, such as demand and supply for specific goods, also dates back to this period. Then, at the end of the 1920s and beginning of the 1930s, the stellar works of John von Neumann and Herman Minkovsky on the mathematical analysis of rather complex models of economic equilibrium were completed. In contrast to physics the methods found were not constructive in nature. Moreover, simplified models of economic equilibrium, like linear input-output models, which could be analyzed essentially with existing methods, were practically never applied because their large dimensions overwhelmed the computational technology of the time. The appearance in the mid-1940s of computers gave a sharp push forward to the use of numerical methods in analyzing models of economic equilibrium. Thanks to this, at the end of the 1940s Wassily Leontief analyzed with numerical reduction a special (but quite important in practical terms) category of equilibrium models, now known in the literature as input-output tables.

Initially independently of computers, then using them at the close of the 1930s and start of the 1940s, because of the remarkable work of Bellman, Dantzig, Kantorovich, and Koopmans, constructive methods for analyzing both optimality and equilibrium models were found that enabled the solution of individual economic problems to be reduced to numerical terms. These methods were already fully comparable with those in corresponding developed physics models. Of course, the level of generality of the phenomena in the physics models is still significantly higher than in economics, where numerical methods of solving complex problems still pertain to relatively special situations.

All that I have said regarding precise modes of economic representation does not exclude the fact that until the appearance of these modes, the end of the nineteenth century, economic theory had been trying to order myriad facts about economic development and find the means to represent them. These generalizations were primarily connected with analytical representations and the effort to derive secondary laws. However, the laws were mainly speculative. This was reminiscent of the situation in medieval physics with its largely speculative principles based on Aristotle.

Introducing secondary laws based on empirical evidence under conditions where the necessary statistical basis is insufficient may lead to very dangerous results. Conclusions about the future and recommendations for action emanating from these laws may prove ineffective and even destructive.

A good example of this predominantly speculative approach to economics was Karl Marx. Above all his concepts are characterized by an effort to derive secondary laws. This relates both to his attempts to develop general laws of human history, as well as his relatively special laws for the economic development of capitalism — despite the fact that not one of these laws has been borne out. The conclusions Marx made concerning the inevitability of a communist socioeconomic transformation were based on his own secondary linear law of the progressive development of social formations from primitive societal structures to communism. But already by virtue of the lack of a sufficient statistical basis for forming such an analytical representation of history (in particular the histories of China and India were poorly known), the laws of history Marx devised were of rather dubious substance. An analogous fate awaited his special laws governing capitalism, like that of absolute and relative impoverishment and diminishing profitability.

Marx did not even pose the problem of seeking modes of representing different socioeconomic systems on the basis of primary laws. He considered such a representation appropriate only for studying Robinson Crusoe or communist society.

I do not wish to belittle the role of research in deriving secondary socioeconomic laws on the basis of experimental data, particularly if these laws concern particular individual aspects of socioeconomic activity. They may, as I indicated earlier, be used not only for extrapolating future development, but also as the basis for building equilibrium-optimality models.

Considering the limitedness of historical experience, however, possibly the center of gravity for research into economic systems ought to be focused on the development of selective methods for representing these systems based on primary laws pertaining to the systems conjugate to them. Moreover, the history of the economic system itself may be studied using selective methods. History may be imitated, but its laws will play an additional verifying and guiding role.

TELELOGY, TELOLOGY, TELEOLOGY

The mode of economic system representation examined above explicitly did not include a time factor, system dynamism. At the same time in hidden form, dynamism in one way or another is present in all of them. Dynamism is connected first and foremost to the system's teleologicalness. Therefore in examining the diverse methods of system representation, the issue of their relation to teleology cannot be skirted.

Taking two extreme cases, we can interpret the source of dynamics of an economic system as being either of random or goal oriented nature; in the latter case it is under the control of some demiurge, for example, a planning state. This approach based on extremes was prevalent for a long time. One example is the antithesis of market versus planning.

Still, there is a multitude of intermediate types of sources of dynamics housed between these two polar ones. Thus, orientation-direction may be sufficient to provide dynamic impulse to the system. At the same time orientation can be supplemented by the category of goal, which represents the final state toward which the system moves. Such a state, not necessarily static, may represent a stationary process.

Furthermore, an absolute goal, even if set in advance, is not necessarily attainable. It may assume the form of an ideal — a state beyond reach. The split between the utopians (in this context, this would include the communists) and proponents of a social welfare system is precisely over this issue. For the utopians the final state is attainable whereas for the latter it represents an ideal. Still, both cases are characterized by the presence of a closure — a state that closes the system and toward which the system strives in its development.

Generally speaking, the goal does not necessarily represent a known quantity, that is, it can be a sought after variable. Therefore, in our subsequent analysis it will be essential to assume that in each particular case the goal is known. This allows us to create new boundary conditions and to close the system from above.

This leads us to the next distinction — that between the goal and the criterion of development. A goal comes in the form of some final state to be attained; in a sense, it represents closures that close the system from above. A criterion indicates certain paths of development leading toward the goal. And because the structure of the dynamic problem, especially if formulated explicitly as an optimality problem, incorporates the initial conditions and the criterion of development, the proposed terms are helpful in differentiating between the category of goal as a closure and the category of optimality criterion as means of selecting the desired course of development.

Distinguishing between the notions of a goal and a criterion renders clarity and rigidity to the terminology used in the formal structure of dynamic problems. Indeed, given the general course of development, one has sufficient information to formulate a dynamic problem. For instance, firm performance can be viewed as being directed toward generating the maximum level of profit. Other goals, such as attaining a certain volume of output, may come into play. Under certain conditions this quantity can be obtained independently based on past extrapolation. Here, a firm's behavior would be directed at minimizing the cost necessary to reach the set goal. Although the firm's overall goal of profit maximization does not change, the distinction between the goal and the criterion employed in choosing the course of development becomes operationally significant.

Other less formal problems can also gain from our classification of the categories of goal and criterion. Take ideologies, for instance. Differences among ideologies are largely rooted in the axioms underlying the choice of the course and goal of development. The profound split between communists and social democrats is that the primary objective of the communists is to attain an absolute goal — creation of a communist society under the leadership of one party, which elaborates and guides the development toward that bright future; for social democrats, as expressed by one of the founders of the ideology, "movement is everything, goal is nothing." In other words, the essential task is to formulate the overall trend aimed at social justice and protection of the weak; here, the actual course of development is elaborated via a democratic political mechanism based on an economic system as a mixture of horizontal and vertical economic mechanisms.

I believe the two terms "telelogical" and "telological" are well qualified to bring the distinction between the notions of directedness and goal into focus. The telelogical approach emphasizes directedness of development, its remoteness — tele — from the present. The telological approach, the term suggested by the Greek word *telos*, meaning the end, assumes the existence of a final state of the development process. The widely used term "teleology" incorporates both dimensions: directedness and goal. In fact, etymologically teleology is based on the Greek spelling of the word "teleos," meaning the end, and also contains the root "tele," which means remote. Teleology presupposes both the directedness of the process and the presence of a goal.

Thus, telelogical and telological approaches differ in their interpretation of the category of goal. The telelogical approach pinpoints the direction of motion without explicitly fixing the goal the system can or must attain; the emphasis of a telological approach is upon reaching the goal. Teleology will be reserved for cases where both the goal and the path of development are strongly emphasized.

The upshot of these considerations is the assumption that the teleological principle reflects the fact that changes taking place in the system are guided; something integrates the development of the system along a given course. As a rule, teleological pertains to the global optimality approach because it explicitly incorporates an integrating feature manifest in the criterion of optimality, which guides the development of the system.

Analytical and equilibrium descriptions are commonly thought of as being nonteleological, for unlike an optimality model they lack a global guiding principle or a goal. These two types of representations are associated with the principle of causality. The latter is primarily characterized by a lack of explicitly manifest guiding forces; analytical representation is a classic example of the said principle. A representation is deemed causal even if the forces (attractors and repulsors) are designated in the system but in a differential form and are merely balanced, as is the case with the equilibrium models.

Still, all these methods involve a sense of direction. Analytically it is expressed as a law; equilibrium representation shapes the direction by defining the forces driving the system. Therefore, it is completely unjustified to oppose teleology with the principle of causality.

As a matter of fact, this opposition is not only in violation of the scientific approach, but was also exploited to fuel an ideological struggle. Teleology and theology were fused together because the idea of God naturally leads one to acknowledge the development of the universe as being directed or even goal-oriented. The principle of causality being devoid of any traces of goal-orientation seemed to adhere to scientific rigidity and was thus opposed to teleology, which is associated with theology. Marxist economic thought witnessed a similar evolution: it linked teleology with the central planning mechanism and the principle of causality with the market.

Teleology may take on a wide range of forms — relative, semirelative, or absolute — depending on the completeness of the conditions restricting system development. When, for example, it is said that an economic system is oriented toward satisfying human needs, this statement emphasizes its teleological character. If it is said, though, that an economic system is seeking to maximize satisfaction of human needs, this stresses the system's optimal shape. Under given conditions, (given resources and technologies), development will attain a certain final state. This state will therefore be relatively teleological. If it is assumed that human needs are finite, that the system is capable of achieving a final state at which human needs will be fully satisfied, it becomes semirelatively teleological.[4] Thus, semirelative teleology is determined by the shape of the function expressing the criterion for which there is an immanent absolute optimum point. If the final state of system development is revealed in an assumption about its limitations only by technologies, regardless of the initial resources and type of optimality criterion, an absolute teleological state is being sought. It is precisely this position that has a place in turnpike theorems.

Analogous considerations may be applied to the equilibrium representation of an economic system where the choice of function linking attractors with primary parameters (the guiding vector function) is an analogue to constructing a scalar optimality criterion (remember that the difference between them is only in the form of representation — differential for equilibrium and integral for optimization). In seeking a condition of homeostasis, we are dealing with relative teleology. In seeking a dynamic equilibrium (or a turnpike) we are dealing with absolute teleology. For semirelativity, equilibrium is defined as the attainment of a state where the values of the guiding vector function equal zero.

Generally speaking, an analytical representation of an economic system can be teleological only in the sense that it reflects only directedness ending in infinity and without a goal. However, an analytical system representation may be semirelatively teleological if the function describing the system is organized so that it leads irreversibly to a final state — a goal. Revealing the goals of analytical representations is quite a powerful operational technique and is used in finding efficient methods for solving differential equations (Bertalanffy, 1968).

Regarding our discussion of telelogy and telology, it is interesting to note the terms applied by Ernst Mayr (1982) to biological systems. Mayr, as is conventional, speaks only of teleology, including in it the goal as well. Meanwhile, he distinguished teleomatic and teleonomic processes. The first one concerns goal-oriented development in systems where there is an explicit choice of behaviors. The second one expresses "any process . . . in which a definite end is reached strictly as a consequence of physical laws" (p. 49). The last approach can be used for analytical purposes in clarifying what eventually may happen if a system continues along past development trends.

It is well known that the market may be approached as an optimization mechanism (Samuelson, 1949) just as a planning system may be approached as an equilibrium-seeking mechanism. To cite examples, the latter approach was notably expounded in relation to the Soviet economy by Victor Danilov-Danil'ian and Michail Zavelsky (1975).

To summarize the preceding discussion about the associations that different economic system representations call forth — because reactive, as well as equilibrium, system representations are associated with market economics, and optimality approaches with planning — these associations are of major political and ideological significance.

Modes of Representations and Politics

From a political standpoint, the variety of ways of representing the economy are not innocuous. Reactive representations are associated with the extrapolation of past trends, and a practice could completely run counter to the current interests of a country's leaders. This circumstance is especially acute for dictators, whose interests to an even greater extent could conflict with past development tendencies in a country or countries with other political structures. Precisely because of this at the end of the 1920s, when Stalin set before the Soviet Union the supertask of creating heavy industry in a minimum timeframe, various reactive economic system representations based on extrapolating from the experience of developed Western countries were ostracized. Formulation of the problem in terms of a global optimization model were in full accord with the dictator's requirements.

The optimization model with an internal global criterion, and not some other optimization model, is generally attractive to a centralized economic system. Thus local optimization models have a quite unpleasant political flavor for centralized political systems, particularly those of an imperial type with heterogeneous regions. The problem is that the regional aspect of economic system representation will, in this event, be explicitly expressed in the interests of each individual region and its potentialities. An explicit synthesizing principle for the system as a whole, which shows how incomes earned by each region are redistributed, must also be introduced. Let us say that in an imperial system, such as the USSR, there are a variety of different regions, some quite rich in natural resources and others poor in them. A local optimization representation could in this case trigger “unhealthy separatist sentiments” in regions with greater natural endowments because they would be able to see clearly how part of their income was being taken away.

Modes of Representations and Ideology

Economic system representations also interact with ideology. Ideology’s influence on economic systems is felt first and foremost in defending or repudiating centralized control — market or state control over the economy — such is the ideological dilemma. Proponents of centralized control, when they begin to represent their system in strict terms, will prefer a global optimization model. Market advocates will prefer one of equilibrium or local optimization.

Thus, for example, the development in the 1960s of mathematical economics in the Soviet Union proceeded principally on the basis of global optimization models. These models were perceived by the powers that be as appropriate for a planned socialist economy and corresponding to a Marxist-Leninist world view. Because for a long time Western economics made use predominantly of analytical and equilibrium representations as well as local optimization ones, Soviet ideologists were convinced that these models were appropriate to a market economy and a product of bourgeois ideology.

An analogous situation occurred in the eighteenth century with respect to physics after the appearance of the extremal principle with the sole difference that ideology’s role was fulfilled by religion. The proponents of equilibrium models thought they were describing a world of cause and effect with no God; the advocates of the extremal principle insisted on God’s creation of a complete world, corresponding to a particular optimality criterion. If one takes into account that at that time the church in Europe was not separate from the state, this kind of ideological interpretation of the world is seen as not innocuous to scientist-atheists. It took a long time before the connection of the two approaches to the study of the physical world was understood mathematically. Then the conflict subsided. The separation of the church from the state ultimately translated the discussion to the realm of scientific thought about the multitude of methods for representing the physical world.

However, even to this day some professors in Catholic universities maintain a heightened interest in extremal principles of mechanics. Moreover, the Catholic Church has exhibited an interest in using active modes of representing economic systems. Evidence of this interest is a two volume set, published by the Papal Academy of Sciences, devoted to a scientific conference on the problems of economic planning. The conference was held from December 7 to 13, 1963. An array of the most visible economists from various countries, all in one way or another involved in the problems of state control of the economy, participated. The Holy Father, himself, appeared before the gathering.

At the same time ideology’s role in forming representations of the economy is quite conflicted. On the one hand, ideology is the breeding ground for various kinds of heuristic notions aiding the appearance of a variety of different representations. It allows the scholar to concentrate on a particular approach and, thus, sufficiently penetrate the essence of a given representation. Thus profound faith in an ideology of centrally planned economics helped Kantorovich at the beginning of the 1940s develop the concept of optimal planning for which he won the Nobel Prize. Two centuries before, Maupertuis’s religiousness inspired him to formulate the extremal principle in mechanics. A scholar’s ideological dedication plays no lesser role in his survival if the country is dominated by a single ideology and if deviation from it may be fatal.[5]

On the other hand, ideology can also act as blinders on a researcher’s eyes. Instead of seeing the commonality of different modes of representation and thinking about combining them, thus deepening his understanding of the approach he has chosen, a scholar’s ideological dedication pushes him to resist certain concepts. The ideological struggle that results may not be the best crucible for scientific progress.

All the same, regardless of the negative aspects of ideology’s influence on the scientist’s formation of concepts, the dialectic of its influence on the development of system representations may force the scientist to disassemble things in order to reconstruct them later.

It would seem that inasmuch as scientists are able to understand the objectivity of a variety of representations and see where it is useful to utilize one or another of them, they may be free from ideological stratification and in-fighting. But, alas, this does not in any way preclude in-fighting among scientists. Fighting does not occur simply because of certain people’s desire to install one mode of representation as the only one. That occurs rarely but shows that scientists do not yet understand the need for a variety of modes of system representation. The more complicated struggle begins over the proportion of different modes to use and the drive to limit the domain of methods already employed when new modes appear. In this instance those scientists who are adherents of a particular mode of representation have to progress, for example, agree to a reduction of additional resources to be allocated to further develop their themes or even to a redistribution of resources overall.[6]

NOTES

1. There is an affinity between the described situation and the category of power in physics. Power denotes the amount of work that can be performed by a

system in a unit of time. Work, in turn, stands for distance times force. Converting work into energy, power indicates the amount of the system's potential energy. The actual realization of this energy assumes various forms.

2. An analogous phenomenon occurs in biological systems. The selectivity in choosing among the various alternatives takes place through the central nervous system endowed with a sophisticated evaluation mechanism (for example, emotions and pain) tied up with reflex mechanisms that limit at times the peripheral nervous system because the latter permit quicker response.

3. I base my discourse on the laws of physics predominantly in the book by E. Rogers (1966).

4. The concept of eschatology, used in theology, characterizes this kind of perfectibility of development, for example, paradise and nirvana.

5. In my work on Leonid Kantorovich (Katsenelinboigen, 1978–1979), I showed that in an autocracy, a nonconformist may survive as a scientist if he at least conforms to the ruling ideology. The combination of nonconformity in science as well as in ideology has a great probability of condemning a scientist to death. The most prosperous scientists conform in both areas. Between the last two groups are conformists in science and nonconformists in relation to the dominant ideology.

6. The history of the Central Institute of Mathematical Economics of the USSR Academy of Sciences may serve as an example of this kind of struggle among scientists to install a variety of modes for representing a socialist economy. In part this history is described in my book (1980).

7

Programming III: Methods of Analysis of Economic Systems

In principle, the discussed modes of representation of an economic system are interwoven with the method of analysis, that is, with the modes of operation of the corresponding system. Our discussion is limited to programming methods of analysis, methods that are proven to be convergent.

We often face the problem of having to construct models of systems having both known and unknown variables. The choice of a particular course of action involves finding the relations between known and unknown variables so that unknown magnitudes can be deduced from the known ones. This task usually involves a two-stage process consisting of the formulation of the problem and the solution of the problem. The borderline separating two stages of behavior organization is rather arbitrary. There are several competing methods for formulating as well as solving a problem. It frequently requires considerable ingenuity to determine which information belongs in the formulation of the problem and which belongs in its solution.

Methods of decomposition are a perfect illustration of this idea. Partitioning a large system into subsystems, with each one being assigned a specific formulated task, and subsequently integrating these subsystems is precisely the method of solving problems of large-scale systems. Thus, the formulation and solution of problems are tightly interconnected. As a rule, however, the framing and the solution of a problem can be separated into two stages of research from the start. Therefore, researchers find degrees of freedom in the modes for representing the system: what to apportion to the problem's formulation and what to its solution.

PROGRAMMING METHODS OF ANALYSIS OF ECONOMIC MODELS UNDER CERTAINTY

The two basic classes of mechanisms, reactive and selective, differ from one another precisely in the way the problem of the economic system's performance is formulated and solved. Thus the reactive mechanism clearly assumes a response based on accumulated experience. Here, the formulation itself dictates the method of solution, meaning the problem is formulated as a law linking inputs and outputs of the system and assumes a solution mechanism in the form of a reflex response. Formally speaking, the problem is reduced to an analytical representation. The inductive approach within this analytical framework hinges on accumulating and processing enough statistical data to arrive at the formula. All the techniques for finding the analytical expression are based on the information pertaining to the given system. Relations reflecting the system's interaction with its environment are not essential.

A system's representation under a search mechanism is a different story. Here, the formulation of the problem as a search for an optimum (equilibrium) implicitly assumes that the general analytical solutions of the problems incorporated into the assumptions of a given problem are given; that is, respective analytical relations pertaining to contiguous problems are made manifest. In the realm of economics, this procedure entails finding the utility function within the sociopsychological system. In the sphere of R&D, it calls for finding the engineering functions linking output with the required resources. It is further assumed that the model linking the two types of analytical representations will generate a selection mechanism designed to select from the large number of possibilities those that satisfy the conditions of the problem. At this point, I would like to examine in greater depth the search methods of solution, particularly with respect to large-scale problems (*Studies in Resource Allocation Processes*, 1977).

Selective methods comprise two broad categories: holistic and decompositional. This is accomplished either by means of local interactions — a horizontal mechanism — or via a coordinating center — a vertical mechanism. Vertical mechanisms assume the subordinate units must obey the orders of the superior one; horizontal mechanisms are predicated upon parity of the interacting objects.

When searching for a solution, holistic procedures encompass the problem in its entirety. This class includes methods of linear programming. The core of these procedures is a rectangular matrix of coefficients of equations (inequalities) of the problem. Because the number of columns of the matrix exceeds the number of rows, the task is to make the matrix square by eliminating an appropriate number of columns subject to certain constraints.

Decompositional methods further split into methods that partition the system into equivalent or inequivalent parts (subsequent aggregation accompanies both). Partitioning a system into disparate parts is typical in solving large-scale problems. This involves discriminating between basic and nonbasic elements (from the standpoint of their functional role) from the set of unknown variables.

Belonging to this class of problems is the analysis of large-scale systems using the method of abstraction-specification. Abstraction procedures are based on isolating basic and nonbasic parameters (Volkonsky, 1966; Ershov, 1966, pp. 249–61). At first, the solution proceeds over the set of basic parameters ignoring nonbasic ones. Eventually, the latter are also taken into account. Many systems, including economy, feature structures that yield to this grouping of parameters. For instance, metallurgy might focus on such essential factors as iron ore, coke, pig iron, electric energy, and labor, distinguishing them from less vital parameters (miscellaneous substances like volframite).

A large class of methods designed to solve large-scale problems is based on subdividing the problem into equivalent aggregates, solving the problem in terms of these aggregates, and then integrating the local solutions. I shall term these methods aggregation-disaggregation. The disaggregation segment is intended to emphasize the unity of the two processes. Economic science has slighted the problem of the synthesis of aggregated parameters at the macrolevel with the disaggregated parameters at the microlevel.

The general solution employing aggregation-disaggregation procedures could be illustrated by respective mathematical methods of solving large-scale systems of linear equations (inequalities). The primary tool of the aggregation-disaggregation methods is to break all the rows and all the columns of the matrix into aggregates, that is collect a certain number of rows and columns in a single group (Dudkin, Rabinovich, and Vakhutinsky, 1987). As the iterative process unfolds, solutions of smaller-scale problems derived from the said aggregates alternate with solutions of problems within each individual aggregate. Structures resulting from the coalition of the two methods (aggregation-disaggregation and abstraction-specification) shall be termed abregates; the corresponding process shall be called abregation.

Methods of decomposition can be classified as parallel, sequential, or para-sequential. Each class is based on vertical or horizontal mechanisms of operation.

A parallel method works by dividing a system into subsystems, which can, by definition, operate in a parallel fashion. An example of this kind of procedure, implemented deterministically via horizontal mechanisms, is the various algorithms used in the collective behavior of automata having no central control (Tsetlin, 1969).

A sequential procedure operates recursively: information generated at step i is used at step $i + 1$, and so on. There are two types of sequential procedures: those that start from the end and those that start at the beginning. The first type includes deterministic methods from dynamic programming elaborated by Richard Bellman (1957) and maximum principle by Lev Pontriagin (1964). Development of the second type of method lags behind the first. Especially interesting in this connection is an attempt by Ron Atkin (1972) to develop deterministic methods for subsystems operating from beginning to end because most methods based on this approach are indeterministic.

Para-sequential methods are exemplified by the decompositional algorithms. All these algorithms, whatever their form or shape, break the problem into the master problem-system and blocks-subsystems, each with its own internal network.

According to the Dantzig-Wolf (1963) decomposition algorithm, each constituent unit is represented within the framework of an optimization scheme as follows: the master program defines parameters (prices) for the unit's optimality criterion as profit maximization. This given, the internal limitations for each unit are taken into account. A unit in turn may be divided into subunits. An optimality criterion of profit maximization is formed for each subunit as well. The internal constraints pertaining to each subunit are thereafter likewise established. Information exchange occurs in the course of this iterative process, allowing each unit to obtain from above the information (prices of the outgoing outputs and the incoming inputs) needed to form its optimality criterion. In its turn, each unit sends from below to the master program information (in the shape of feasible plans) about its own capacities.

In this entire procedure the system's representation does not have to be isomorphic with respect to the subsystem's. The global optimality criterion for the system as a whole in no way has to be isomorphic to that of the subsystems: the latter may be invariant relative to the form of the global criterion. Misunderstanding of this kind of interconnection between the economic system representation and its constituent subsystems leads to major theoretical and practical miscalculations.

Thus, for example, in the former Soviet Union (as well as other socialist countries) many economists are convinced that under socialism profit cannot be the prime criterion for enterprises because development for the system as a whole is supposed to proceed according to the basic economic law of socialism, which places priority on maximal satisfaction of the continually increasing needs of society's members. Things are different, they reason, under capitalism. There, because of private property and the resulting indirect expression of social interests, each firm pursues the goal of profit maximization.

Naturally, if society is pursuing particular objectives requiring its enterprises to observe additional conditions besides profit maximization, these may be formulated as additional constraints. In other words, profit in this instance may be an insufficient condition to describe enterprise behavior and must be supplemented by other indicators. Posing the question in this fashion is different from simply denying profit any role as a leading indicator in evaluating enterprise behavior merely because it is insufficient. Even in capitalist countries, where profit is systematically used as an indicator of a firm's performance, the drive to increase it may be curbed by generally accepted societal interests (for example, firms on government contracts are obligated to hire a minimum percentage of minorities).

Another instance illustrating the misunderstanding of the connection between the modes of system and subsystem representation, which I had occasion to encounter, concerned the creation of an automated control system for an oil refinery. The refinery was represented as a hierarchy, and the Dantzig-Wolf decompositional procedure mentioned above was chosen as its control algorithm. The engineers designing the refinery were interested in this algorithm's economic interpretation because they were hoping the ideas contained therein could help modify it. Because the algorithm slowly converges, the need to modify it for facilities with rapidly moving processes working in real time was quite persistent.

The engineers quickly grasped that this algorithm has a serious content and is not simply a collection of mathematical manipulations. George Dantzig (1963) provided an excellent explanation of the content of this algorithm that greatly eases its understanding.

It was difficult, however, for the engineers to understand what happened next. For some reasons unclear to me, the project's client demanded that the enterprise be operated on the basis of maximizing the output of a given product mix. And the engineers did not understand why this criterion could not just be depicted in the problem for each individual part of the enterprise (its components), why the individual units needed a criterion based on profit maximization.

PROGRAMMING METHODS OF ANALYSIS OF ECONOMIC MODELS UNDER UNCERTAINTY

A vast amount of literature is devoted to the problem of performance of an economic system under the conditions of quantitative uncertainty (especially so when uncertainty is presented in terms of probabilities). The above considerations regarding the methods of formulation and solution of problems were rather general. I merely illustrated them using extreme cases manifesting certainty of quantitative information and completeness of qualitative information (this corresponds to cell 11 in Figure 7.1). These considerations apply equally well to tasks characterized by incomplete information regarding the system, including the construction of a model of the system as well as algorithms to analyze its performance. Of course, the assumption of the convergence of these algorithms is still valid.

The above case also accommodates methods that allow for prices to be set as the process of search for the equilibrium/optimal state unfolds. Here, prices also convey complete condensed information about the environment, but only in terms of probabilities (Dynkin, 1971). In the general case of reactive methods, the initial state as well as the resulting final state is given in probabilities terms. Finding correlations is a typical example.

My focus shall once more be placed on the selective methods. First, I shall consider situations of quantitative uncertainty under complete qualitative certainty. In this context, what makes selective methods different is the presence of uncertainty both in the formulation and in the solution of the problem. Working under a simplifying assumption that the degree of uncertainty is given by two extreme states — certainty and uncertainty — we arrive at a 4-cell matrix (see Figure 7.1).

There are numerous algorithms to solve problems resembling cell 12 of Figure 7.1 (the problem is formulated in terms of certainty but the solution allows for uncertainty). Here, another relevant aspect of the problem is to identify it with either the equivalent or inequivalent methods of decomposition.

The said methods for solving large-scale problems divided into nonequivalent parts best fit situations that yield basic elements that call for highly sophisticated and precise techniques and nonbasic elements whose value can be computed using less intricate random techniques. One example of this algorithm executed via a vertical mechanism was elaborated by Kantorovich in connection with the "knapsack" problem. The problem is to fit a large number of different objects into a knapsack. First, Kantorovich roughly divided all the objects into large ones and small ones. The solution stipulates using exact methods of linear programming to determine the optimal allocation of large objects. Then, smaller objects are simply "thrown" into the sack, the sack is given a good shake, and the smaller objects are fit in at random (Kantorovich & Gorstko, 1972).

FIGURE 7.1
Combinations of Certainty and Uncertainty in Formulation and Solution of a Problem

Problem formulation	Problem solving	
	Certainty	Uncertainty
Certainty	11	12
Uncertainty	21	22

Examples of probabilistic methods (with the system divided into equivalent parts) executed by means of horizontal mechanisms are methods elaborated by Prof. M. Rabin in the area of computer design. His models based on parallel processing are geared toward solving problems known as the problem of the "dining philosophers" or "Byzantine Generals" (Kolata, 1984).

Let us have a brief overview of methods that belong to cells 21 and 22 of Figure 7.1. They may be united by the idea of operating the system according to observed data, which cleverly combines *a priori* conceptions about quantitative value probability parameters with their *a posteriori* elaborations in the course of performance (Wagner, 1975). In the simplest case, the solution of an economic problem lacking in its formulation complete information can be considered analogous to the solution of a problem formulated under the assumption of complete certainty by algorithms whose very structure precludes probabilistic elements. This case belongs to cell 21.

In the simplest case of a problem formulated in terms of probabilities, each step of the procedure generates the best (in terms of probabilities) solution based on the information available at that step. A system's adaptation to new conditions will take place depending upon the state a system has attained. Moreover, it is even possible in advance to compose plans that differ in their initial conditions. Depending on the realization of a given plan, it is either continued, modified, or replaced by another plan. This represents the so-called contingency planning.

However, under the conditions of uncertainty, this method of control is fraught with considerable losses, especially pronounced when the system has multiple ingredients (Petrakov & Rotar, 1985). Then it is desirable to seek a condition under which the system will be less dependent on inferior implementations of any of its constituent ingredients. Otherwise the remaining components will be significantly underutilized. In order to operationalize this kind of procedure, a new institution must be introduced into the system — reserves. The size of the reserves — complete with their corresponding needs for replenishment, maintenance, and so forth — and their structure (whether in the form of finished goods or the means to produce the same) are unknown variables.[1]

In operating an economic system under conditions of quantitative uncertainty, the use of both reserves and contingency planning tend to support each other.

One important variant of a program-like class of methods is the robustness principle introduced by Jonathan Rosenhead, Martin Elton, and Shiv Gupta (1972). It pertains to cases where information regarding the various strategies is available but the implementation takes place under uncertainty. This principle embodies the idea of flexible structures.

> A class of problems for which (profit-maximizing) optimality is particularly inappropriate is that of strategic planning, a special case of decision-making under uncertainty. A plan consists of a number of currently preferred future decisions, and only one decision or group of decisions which must be irrevocably implemented at this stage. The subsequent preferred decisions which constitute the plan can each be reviewed in the light of up-to-date information when their commitment times arrive. In this situation an appropriate criterion for the initial decision is one based on the degree of useful flexibility for future decisions which will be maintained.
>
> *Robustness* is such a criterion of flexibility in achieving near-optimal solutions in conditions of uncertainty (p. 428).[2]

As formulated above, the robustness concept is general and plausible enough. Its actual application, however, demands a particular class of methods adapted to certain structures or, more precisely, to the conditions of quantitative uncertainty when we have a set of clearly defined sequential tactics and need a strategy to schedule these tactics in a sequence ensuring the optimal overall trajectory.

Systems operating under qualitative uncertainty seem to be based on the method of aggregation-disaggregation. Methods of abstraction-specification of input and output parameters can be used to supplement the said method. However, the latter are intrinsic to the situations of qualitative uncertainty, when various parameters are dropped from consideration (or are perhaps unknown).

Abregates thus generated can be used to combat the dimension curse mentioned above. Another area of their application is situations where many new objects that satisfy certain requisite conditions can appear within an abregate. One prevalent feature of this class of situations is the need to link the design of new products with the products' subsequent production. Here, the issue is not so much the design of the new product as the system's preparedness to assimilate it when its novel qualitative features become apparent. The following example relating to the production of shoes illustrates this feature. The shoe industry produces a great variety of footwear. We know that new kinds of shoes will appear in the future although we do not yet know their concrete styling. In order to establish connections between the shoe industry and both its suppliers and consumers, it is sufficient to know about shoe production in terms of abregates. At the same time, in order to implement fully and concretely the system's operation (from supplier to producer to consumer), conditions have to be created under which these abregates will be concretized in accordance with specific knowledge about the desired shoe types.

This sort of a process may be based on the definition of a basis ingredient for all abstractable objects. This ingredient must be one not only inherent to all these objects, but also something that to a large extent determines the demand for all other ingredients required in the manufacturing process.

In accordance with this principle, we could create an abstract called "leather footwear," inasmuch as leather is the versatile material used for all types of shoes in the given abstract and the manufacturing of shoes from leather presupposes the utilization of a corresponding amount of equipment, labor, and other factors. However, this abstract takes no account of the color of the leather, type of buckle, or other details. This abstract predicates an aggregate corresponding to the respective averaged-out indicators of expenditures on the various factors of production.

The next task is to specify shoe-type and disaggregate the expenditures. This requires that the process of abstraction and disaggregation be carried out in conjunction with finding versatile equipment that could be easily converted from the production of one type of shoe to another. Meanwhile, new shoe fashions may require one-time modifications in means of production, for instance, the need for new lasts. Therefore, as a precondition to the process of specification-disaggregation, the potential need to quickly fabricate new lasts must be foreseen. This could be done either by organizing specialized shops for producing these directly at the shoe plants themselves or through specialized last producing companies. In all cases the versatile inputs (wood or aluminum, plus the corresponding equipment and labor) to assemble the new lasts are necessary.

Under different kinds of economic mechanisms, the abstraction-aggregation process will take place in different ways. Within the framework of a market economy, the aggregated indicators will be used to establish the relatively long-term relations between participants in the shoe system. Special subcontractual or other methods will be used to govern the exchange of concrete goods over comparatively short time periods. Within the framework of a planned system the same objective is accomplished by the higher organs establishing plans for the lower ones in terms of aggregated indicators and by the latter concluding direct contracts between themselves to concretize the specifications and terms of supply (Katsenelinboigen, 1988).

The examined methods of operation under the conditions of qualitative and quantitative uncertainty are not divided by a Great Wall of China. These methods are deeply intertwined, something that becomes manifest under quantitative uncertainty in the creation of reserves and in the creation of versatile means as preconditions of operation under qualitative uncertainty. Because versatile means represent an effective way to maintain reserves, both the reserves as well as the implementation of abregates may require the same versatile means. For instance, it may be more advisable to keep in reserve at the assembly line not the entire set of the required components but only versatile equipment and materials from which any of the known components can be manufactured, provided the need for these components varies probabilistically. These versatile means can be employed to produce new components for new products that were conceived when the abregate was formed but that were naturally unknown in all their minute details.

NOTES

1. It is interesting to note that, under otherwise equal conditions, as the probability of obtaining the necessary ingredients increases, the living organism keeps smaller reserves of them on hand: without air a person can survive only a few minutes, without liquids several days, and without food several weeks.

2. For a detailed explanation of the concept of robustness, see Rosenhead (1980).

III

POTENTIATING

8 Generalities

My greatest strides belong in the realm of alternative elaboration within the framework of a potentiating method, which constitutes the core of this book. At this point, I call the reader's attention to one key section in this chapter entitled "Methods of Integration of Parameters Comprising Potential," which presents a cohesive and concise overview of the subject matter suggested by the title. This section is important in view of my rather atomized discussion of various works dealing with specific methods of integration of parameters comprising the potential. This loose organization of the material reflects the logic of hierarchy of economic systems rather than methods of integration of a potential's parameters.

GENERAL REMARKS

The potentiating method can be interpreted in a broad or in a narrow sense of the word. Broadly defined, it entails procedures designed to solve certain multistage problems, that is, problems having multiple transitions. The multiplicity of transitions may arise simply as a result of a large number of participants, with the behavior of a given participant at a given moment of time depending on the planned actions and the state achieved by some other participant.

In a more narrow sense of the word, the potentiating method involves one single stage of this multistage process, with the problem of interlinking all the stages left unaddressed.

The present book focuses on the second perspective. Only one chapter, Chapter 16, examines the various ways of incorporating the proposed constructs of evaluation of a given stage within the overall solution scheme. This chapter analyzes the various styles of leadership in the firm. Sufficient formalization of the proposed method of potentiating (in the narrow sense) makes quite plausible its inclusion into a number of well-known algorithms that are compatible with certain methods of evaluation of individual stages. One such algorithm is the so-called Genetic algorithm proposed by John Holland. This algorithm (Goldberg, 1988) assumes that the simulation procedure incorporates objective function values denoting such basic structure as an individual string, an encoded set of a system's natural parameters. The structure of this objective function may vary. It could be represented by profit or by utility or by some beneficial properties. The weight function examined in this book fulfills all the requirements imposed upon the objective function in question.

I summon systems approach to examine the proposed principle of the creation of this weight function. In terms of the functional point of view, the principle essentially aims to augment the potential at a given stage, that is, to increase its impact upon the development of the system directing it along the course selected based on one's beliefs. Thus, from the functional point of view, uncovering predispositions means finding a state that will ensure that the system is prepared to face an unknown future.

From the structural point of view, the principle is designed to distinguish certain features that allow the current state of the system to be atomized while the holistic effect that reflects the influence of the given state upon subsequent development is preserved.

From the standpoint of the process, the search for predispositions entails a synthesis of elements characterizing the state of the system. In fact, the true task of elaborating such an intricate structure of elements is to integrate the system, to interface its various subsystems (both in time and in space) and, at the same time, preserve the holistic effect.

There are two modes to form predispositions and preserve this holistic effect: reactive and selective. The first method is based on a set of actions that represents isolated rules, heuristics. The second method involves a search for a scalar value denoting the system's predisposition. The cons and the pros of the two methods will become clear after reading other parts of the book. My subsequent discussion will spotlight the second approach, namely the search for a scalar value reflecting the degree of predisposition.

From an operatorial perspective, the creation of predispositions hinges not only upon the power of the operator, but also upon certain operator-specific characteristics, stipulated by its design or formed over the course of its operation. An operator's subjectivity (one reason) stems from the fact that the implementation of the potentiating method depends on an operator's confidence in this ability to realize the potential generated at a given stage, to find ways to drastically augment the sytropy of the system in an indeterminate future.

In terms of genesis, the search for predispositions entails finding the sources, which form the structural components comprising the state of a system.

TYPOLOGY OF PARAMETERS REPRESENTING THE STATE OF A SYSTEM UNDER POTENTIATING

The structural aspect of a systemic search for predispositions merits a closer look. Our first task in analyzing the structural dimension is to develop a typology of parameters that represent a system under the potentiating approach. Taking the general case (see Chapter 5), representation of a system assumes that its state is characterized by a pair of variables: primary ones and conjugate ones (the latest could be attractors and repulsers).[1]

As soon as the system becomes more indeterministic the primary parameters split into essential and relational ones. Essential parameters form the skeleton of the system; another name for them is material parameters. Resources and products belong to this class of parameters. Relational parameters denote the various relations among the essential parameters.

Relations can be divided into the following subclasses:

Attributes signify links between the object and its environment as an entity; this includes parameters of time, coordinates, and color.

Structural parameters denote the ratios between the primary parameters, such as the ratio between the volume of production and the utilized labor.

Characteristic parameters are the features arising in the process of interaction among objects (for example, substitutionality of objects, compatibility of objects, and universality, the relative location of objects (proximity, centralization, or openness), and the properties of actions performed by or over objects (monotonousness or traditionality).

Table 8.1 presents a list of the most common characteristic parameters with examples borrowed from different domains of economics: the stock market in the realm of macroeconomics, strategic minerals supply at the branch level, and corporate planning at the microeconomic level.[2]

METHODS OF INTEGRATION OF PARAMETERS COMPRISING POTENTIAL

One feature shared by both the deterministic models of economics, such as the Arrow-Debreu model, and the indeterministic model is the separateness of a productive unit (subsystem) and the consumer. Consumers are assumed to have their subjective utility function, which is largely determined by

TABLE 8.1
Primary Relational Parameters

Parameter	*Financial Markets*	*Strategic Minerals*	*Corporation*
1. Simultaneity/Successiveness	Do arbitrable opportunities exist due to the inefficient discounting of new information (e.g., dissemination of inside information or price adjustment between different markets for same commodity)?	In an emergency, is it possible to speed the development of more material and processing capacity?	Is it possible to organize the planning process for all subdivisions such that it takes place at the same time?
2. Divergence/Convergence	Is there a tendency in the market to approach an equilibrium price (e.g., existence of bands of trading levels)?	Are constraints on the focal system burdensome so that reasonable goals cannot be met?	Will each subdivision truly work together to produce a comprehensive plan?
3. Mutuality/Preponderance	Does a mismatch between buyers and sellers exist (e.g., market crash or underpriced new issue of stock)?	Does one subsystem dominate another, or is there mutuality and balanced dependence?	Will a single organization prevail in accomplishing its planning goals at the expense of others?
4. Explorability/Tradition	Can new functions or customers be found for security (e.g., long-term bonds used for swap market)?	Are new or revised processes and materials investigated and adopted?	Can we successfully deviate from the accepted way of planning?
5. Sensitivity/Irresponsiveness	How greatly is the price of this security affected by new information from a certain source (e.g., airline stocks and the price of oil)?	Does the production of valued outputs rely on the rate of production or other subsystems, or does it vary independently?	Is the planning process responsive to environmental and organizational information changes?
6. Repetitiveness/ Monotonousness	Does the return from this security grow at a compound rate or does it vary cyclically (e.g., dividends vs. principal value)?	Is the use of materials growing (declining) or fluctuating?	Will planning be a continuous process as envisioned or will there be periodic bouts of planning activity, for example, quarterly?
7. Jumpiness/Smoothness	Do interruptions in trading create irregular price adjustments (e.g., opening price vs. closing price on previous day)?	Is any subsystem susceptible to rapid uncontrollable change that would affect the focal subsystem adversely?	Is a constant level of attention paid to the planning process by those involved?
8. Mobility/Stationary	How volatile is the return of the security (e.g., gold stock vs. food stock)?	Once environmental changes are recognized, is the response completed promptly, or do old habits linger?	Will the plan be directed at preserving the existing order or changing it?
9. Substitutability/ Irreplaceability	Can the security be traded in a market (e.g., commercial loans vs. federally subsidized mortgages)?	Can value outputs be substituted for or dispensed with for a time?	Are there alternative plans?
10. Reversibility/Irreversibility	Can a security be purchased at a prior price (e.g., option contract)?	Can materials or processes be changed and then go back to a previous condition; that is, are there recovery and recycling?	If the new planning process is found to be failing, can we easily revert to the old approach?
11. Proximity/Remoteness	Does the locality of the asset or of trading imply different cost structures for different participants (e.g., foreign loan or foreign stock exchange)?	Are materials and processing locations distant?	How similar are the plans elaborated by various subdivisions?

TABLE 8.1, continued

Parameter	*Financial Markets*	*Strategic Minerals*	*Corporation*
12. Contiguity/Angularity	Do different securities share risk factors, or are they tied only to the value of the underlying asset (e.g., company's common and preferred stock vs. bond)?	Does the supply of a given mineral depend on many factors or one decisive?	Are the values and perspectives of the departments held in common?
13. Mediality/Laterality	Is the price of the security skewed in comparison to its competitors?	Are materials and processes in hard-to-reach areas, or are they near protected, major transfer points?	Will the plans coincide with the central theme of the department?
14. Tributary/Distributory	Are the stocks distributed among several people or concentrated in one hand?	Are the valued outputs used in many sectors equally, or are the uses concentrated in a few, key sectors?	Are the planning powers distributed among many or few employees?
15. Versatility/One Sidedness	How liquid or adaptable is the security (e.g., Treasury Bond vs. Rockefeller Center)?	Is the given mineral used in manufacturing of many products or just one?	What extent of specialization does the plan entail?
16. Saturation/Exhaustibility	Are the security's underlying assets fully utilized (e.g., firms with large cash reserves)?	How many resources and reserves are available for eventual development? How large are intermediate stocks in the supply chain?	Do we plan for 100 percent productivity of resources or are there reserves specifically protected?
17. Reflectivity/Irreflectivity	Does the market's vision of the security's price affect the traded value of the security (e.g., market's boom/bust psychology)?	In an emergency, is it possible to suspend constraints for a time?	In composing a plan, does a given department consider the way other departments perceive its actions?
18. Complementarity/Self-sufficiency	Is the asset more valuable alone or combined with others (e.g., merger consideration)?	Are the outputs combined with other similar things, making them both more valuable, or are the outputs used independently?	Do the individual plans of the subdivisions generate a holistic effect?
19. Cohesiveness/Porosity	How is the asset positioned to fill the market's needs (e.g., niche strategy vs. blanket strategy)?	Are all minerals supplied or are some missing?	Is the planning process complete or does it have gaps?
20. Compatibility/Incompatibility	Do the underlying assets produce excess value by their investigation (e.g., divestiture consideration)?	Are subsystems in conflict with each other?	Are tactical resolutions in conflict with overall plan?
21. Historicism/Markoverness	Does the history of market changes influence the actions of market participants (e.g., short- vs. long-term horizon)?	Is there constancy and reliability in the processes, or can stages suddenly drop out or change?	What is the impact of the firm's history on the creation of new plans?
22. Maturity/Rawness	Is information regarding the potential value of the asset fully disseminated in the market (e.g., failures in capital market)?	Is the infrastructure and processing capacity large and well developed, or are there bottlenecks?	Is the firm versed in the realm of planning?
23. Monopolization/Influenceless	Does the asset derive excess value from its influence over competitors (e.g., block traders vs. individual traders)?	Do a few or one subsystem have exclusive possession or control of materials or processes?	Has some single subdivision monopolized planning?
24. Permeability/Closeness	Do restrictions in the capital markets close potential investment avenues (e.g., private placement vs. initial public offerings)?	Is exchange between subsystems restricted or selective?	Is the plan amenable to external changes?

TABLE 8.1, continued

Parameter	*Financial Markets*	*Strategic Minerals*	*Corporation*
25. Excreticity/Retentiveness	Are there exogenous restrictions upon production and output (e.g., government regulation)?	Does the country export strategic minerals?	Does the information regarding the plan trickle outside the department?
26. Diversity/Uniformity	Does the degree of diversification of investment projects add excess value to the asset (e.g., economies of scope)?	Does the quality of mineral resources vary, or is it the same?	Do different departments employ different methods of planning?
27. Orderliness/Disorderliness	Are there program-like mechanisms to price the asset?	Are there government agencies to determine requirements of strategic minerals?	How ordered is the process of planning?

Source: Compiled by the author.

factors exogenous to the economy proper (psychology or social relationships).

Here, the productive unit, be it an enterprise, a region, or a country, seems to be devoid of its own subjective interests. Its task is to produce output for the outside and consume resources from the outside within the internally defined constraints, including the human factor (such as leaders' interests), while creating potential, which would eventually lead to greater utility for the consumer. This is a strong assumption indeed because the leaders of the country or of a productive unit possess their own aspirations for the subordinated system (subsystem). For the sake of simplicity, let us take this assumption for granted with one mild qualification: ideally it is possible to devise an incentive mechanism for the leaders and the groups such that the interests of both meet asymptotically. In Chapter 14, this issue is discussed in depth as it pertains to the firm.

The description of a productive unit's potential within an indeterministic framework entails one quite plausible assumption — that each such unit possesses its own specific fixed line of output produced and resources consumed. In addition, the set of relational parameters is made uniform because in this particular case these parameters perform like elements of the mechanism of integration as a whole. Under the above assumptions, there are two basic methods of integration of parameters comprising the potential — reactive and selective.

The reactive method performs parameter integration based on a set of independent rules, heuristics, which postulate the significance of single material as well as relational parameters. Chapter 15 includes an entire section dealing specifically with the question of heuristics.

Under the selective procedure, parameter integration is accomplished via the weight function, which, in the ideal case, commeasures the values of individual parameters comprising the potential. The focus of my subsequent discussion is on the selective method of parameter integration via the weight function.

In case of a linear approximation, the weight function must incorporate uniform global valuations for every material as well as relational parameter. These valuations can be generated by means of coordinating bodies, such as consulting firms or research centers, or via the local interaction of the participants, that is, directly via the market mechanism.

By analogy with the deterministic systems, these uniform global valuations of parameters comprising potential should also be marginal, a point that comes to the fore when these valuations are set via the coordinating centers. These bodies have superior capability to set uniform global valuations based either on statistics reflecting the performance of individual production units or on expert judgment. However, this advantage frequently leads to averaged-out global valuations rather than marginal ones; the former are valuable only when used to complement, not replace, the latter. The two approaches used to set averaged-out valuations are illustrated using the PIMS model (statistical approach) and the concept of the potential in the realm of strategic minerals discussed in the expert approach.

In my subsequent analysis, an attempt is made to relax the aforementioned assumption that the leaders' ability to manage can be fully known. The inadequacy of this assumption surfaces under indeterministic conditions because the individuals' creative abilities, which so far have eluded a clear-cut definition, are of major relevance under indeterminism. The infeasibility of incorporating the leaders' creative abilities into the economic model introduces major changes in the way the weight function measuring production potential should be set up. First of all, the set of relational parameters incorporated into the weight function would vary depending upon the creative abilities of the leaders to define the subjective meaning of production potential. Second, when the valuations of parameters of the weight function denoting a given definition of production potential are set, a certain subset of these valuations — primarily the ones pertaining to material parameters — can be made objective. The valuations assigned to relational parameters are largely subjective. In an ideal case, we could eventually arrive at uniform global valuations of relational parameters provided they reflect the full diversity of creative abilities of the leaders.

The present chapter presents a general overview of certain aspects of setting objective global valuations of material parameters. In Chapter 9 the problem is investigated in depth and illustrated with a concrete example of turnpike valuations and their applicability at a macroeconomic level. In discussing the aesthetic method further in this chapter, I shall touch upon the role of the subjective factor in setting the valuations of relational parameters.

THE STRUCTURE OF THE WEIGHT FUNCTION UNDER THE POTENTIATING METHOD

Both essential and relational parameters are associated with respective conjugate parameters. For instance, prices represent a conjugate parameter of such essentials as product and resource; the value of the firm's relative market share is a conjugate parameter assigned to a relation (in this particular case, a structural relation).

This raises the following question: "Why separate primary parameters into essential ones and relational ones?" In a deterministic situation, the local representation of a system, that is, its slice-state, and its link to the long-term dynamics of the process can be

represented by means of essential parameters and their respective prices. Here, the prices are complete, meaning they convey full information to the object as to its proper course of development. In other words, this local characteristic of a state is sufficient, given the discovered priorities (prices), to find the proper way to reduce certain expenditures or increase the output of certain products. However, this approach is adequate when all the elements of the system are interlinked in a complete and consistent manner.

But the existence of gaps in the system that cannot be remedied by this kind of representation of intermediate stages necessitates a different approach. We have to settle for incomplete prices and try to supplement them with explicitly defined relations. These relations are generated as intermediate structures in the course of uncovering complete prices.

This is precisely the reason why relations emerge as independent variables on par with material parameters. The degree of unconditionality of the prices varies according to the degree of indeterminacy of the situation; under total indeterminism these unconditional valuations can even be indistinguishable. In a fully deterministic case, they degenerate into conditional ones. Thus, the valuations range over a broad spectrum depending on the situation.

The described principle of measuring predispositions is based upon a holistic search for the system's potential via its individual essential parameters and relational parameters.[3] What is new about this approach is that it introduces, on the one hand, unconditional valuations for essential parameters and, on the other, relational parameters as independent variables. It is precisely the incompleteness or inconsistency of the links between the present and the future states that necessitates bringing in some unconditional valuations of essential parameters and augmenting them with appropriate valuations of relational parameters. In this sense the relational parameters also become guiding parameters. If I correctly understood the corresponding tenet of Eastern philosophy, one may treat these relational parameters as "causes without energy."

In other words, the creation of predispositions introduces a new class of disjoint parameters, which can assume two forms: a new kind of unconditional parameter conjugate to the existing primary variables consisting of material parameters or a completely new kind of both primary and conjugate independent variables denoting the relations. At the same time, the described approach to com-puting predispositions fails to alleviate incompleteness and incon-sistency. It is the infinite number of relational parameters that is the primary cause of incompleteness in the representation of the system's potential. The independence of the valuations associated with each relational parameter (or a combination of a small number of them) is bound to lead to contradictions in the procedure (due to the impossibility of finding valuations that would account for the overall complexity of the situation).

These disjoint parameters which present an incomplete and sometimes even an inconsistent picture, precipitate the indeterministic essence of the methods that utilize them. First of all, it concerns relational parameters. From the above examples we can see that these relational parameters have been based on nothing more than unsystematic empirical observations. I know of no deductive theory for generating them. Because we do not know if the available relational parameters are sufficient to produce an estimate, a good metamodel should suggest ways of generating new parameters of this kind. A more traditional task of metamodels is to help make the existing value of relational parameters more precise. In this sense, it resembles a common learning model, according to which a metamodel is used for making adjustments in the behavioral model as new data accumulate. Because of their completeness and consistency, deterministic methods can condense global information (for instance, find the Lagrange multipliers), thus encompassing the system using fewer variables, that is, with relational parameters (and their respective valuations) left out.[4]

Perhaps, the need to separate all the relevant parameters under potentiating will become clearer if we recall that the calculation of predispositions is geared directly toward measuring the impact of unique situations upon the development of the system. This calculation entails a unique amalgam of information about the future, as well as past, states of the system.

Typically, unique situations make it impossible to predict the impact of a given step upon the future even in terms of probabilities because it is impossible to reduce these situations to ordinary cases that are statistically verifiable. Generally speaking, a unique situation embodies new essential parameters. Evaluation of these parameters would yield a measure of their impact upon future development. However, the very definition of uniqueness rules out the possibility of deriving these valuations of new essential parameters from the past, simply because these parameters were nonexistent. The valuations of new parameters can only be taken from the future, but the future course of events is unknown for the following reason: there exist no program-like methods, which would yield conjugate variables matching new essential parameters while taking into account the various boundary conditions, such as the initial resources, the feasible objectives, and the feasible methods of transformation of inputs into outputs. Consequently, we have to settle for partial, derived from the future valuations of new essential parameters. One way to find these valuations is based on the information, which is generated by the process of transformation of inputs and outputs, about the new parameters.

However, the found valuations are not sufficient to measure the impact of a given state of a system upon the future flow of events because the interaction of different parameters underlying these valuations is ignored. It is essential to supplement these valuations by other parameters designating the relations among the essential parameters. These relations were implicitly concealed in the conjugate variables (complete valuations) arising under program-like methods of transformation of inputs into outputs. However, under potentiating, it is necessary to stipulate these relational parameters, explicitly and individually, and then establish functional connections among them. The set of relations and their respective valuations (weights) can probably be formed based on previous experience. The feasibility of this procedure is corroborated by the relative stability of the valuations of relational parameters (relations). Just take an example of social systems where the essential parameters, particularly new technologies, which represent them, have undergone tremendous change. At the same time, the relations among people have changed relatively little. One of the major conflicts of modern society is this incongruity between the new military technology capable of destroying mankind and human relations, with strong elements of aggression, which have remained at the level of the long-gone past.[5]

Further formalization of multiparameteric representation of the state of a system brings us (in the simplest case) to a weight function in the form of a linear polynomial:

$$u = sp + \rho\pi$$

The function incorporates four groups of parameters:

s = vector of essential (material) parameters, each characterized by a set of indicators that differentiates the object (make it discrete) within the universe as a continuum;
p = vector of unconditional valuations (weights) assigned to each essential parameter;
ρ = vector of independent (controlled) variables denoted by the parameters of relations among the essential parameters; and
π = vector of valuations (weights) of relational parameters.

Linearity of the weight function implanted in the method of potentiating is not intrinsic to this polynomial. The nature of the weight function could be nonlinear, but the linear case merits special attention. Essentially, mere feasibility of linear representation addresses in a formal way the problem of representing the holistic effect by analytical means, that is, by dividing the whole into constituent parts and then adding these parts. Because of this highly nontrivial carving procedure, which incorporates such novel elements as unconditional valuations of essential parameters and relational parameters treated as independent variables, it becomes possible to represent the whole via the sum of its parts.[6]

HEDONIC PRICES

Before we embark upon a detailed discussion of the described methods of measuring predispositions, I will devote a small section to the problem of hedonic prices. The way these prices are set is frequently confused with the way the valuations of various parameters are found when computing predispositions.

Goods are differentiated by a panoply of properties such as color, size, and weight as well as by other features that reflect the direct bearing of a given product upon the specifications imposed upon other products required for its preservation or operation (for instance, gas mileage of an automobile). Generally speaking, prices for all varieties of goods can be established starting from the general model of equilibrium (optimization). However, implementing such a model is practically impossible not only because of the weakness of our study of consumer preferences, but also because of the model's enormous dimensions. There might be thousands or even tens of thousands of varieties of the same product.

The phenomenon of the dimension curse that arises in connection with this has very important theoretical and practical implications. The struggle against the dimension curse requires the introduction of generalized products and the establishment of corresponding prices. Together with this, product individualization is also required, because without individualization, the benefits of forming generalized products cannot be realized.[7] This requirement to formalize the problem of economic performance is, generally speaking, sufficiently universal and may be attributed to any large-scale system. The problem of creating generalized products and their individualization is formally solved via abstraction-concretization procedures and/or aggregation-disaggregation.

However, the execution of this procedure is complicated by the need to encompass the entire system. Practically, it is still very important to find economic means of setting prices of goods having new features; for instance, this procedure is employed in computing the price index that reflects the real growth of the standard of living when better-quality (modified) products appear on the market. It thus appears natural to quest for procedures of price setting applicable to an extensive assortment of different commodities — procedures that yield disjoint (by each type of product) prices that reflect the particular features of the product.

The process of setting prices on goods with their qualitative peculiarities taken into account can be solved under conditions of an extensive product diversity in two ways. The expounded methods are characteristic of creative acts in general, and they may be more transparently demonstrated by an example from the world of sculpting. One technique is based on taking the whole object, a stone perhaps, and cutting out all that is unnecessary from the point of view of the sculptor's goal. The other method involves gathering as a foundation some part of the future sculpture and combining it with a number of different parts, which in the final synthesis will yield the desired image.

To a certain degree this analogy is pertinent in constructing prices on individual products. The first method mimics the process of search for an equilibrium, the process incorporating all kinds of generic products, that is, products with their potential features. Subsequent individualization of the actual products proceeds by gradually shaving the generic product of all the features that a given product lacks. The second technique is based on the statistical method of factor analysis of the qualities of a particular good and subsequent modification of these qualities, using as a basis another good selected as a benchmark.

Hendrik Houthakker's (1952) pioneering work proposed a methodology for finding prices for varieties of goods. It was based on statistical study of the different qualities of the goods as independent variables in conjunction with the independent variables denoting a good's quantity.

At the beginning of the 1960s, Zvi Griliches undertook to enliven the aforementioned approach to studying price indexing by using multiple correlation methods. A large group of U.S. economists, intrigued by this work, conducted significant research into establishing prices on varieties of several kinds of goods (*Price Indexes and Quality Change*, 1971). These prices were termed "hedonic."

To illustrate hedonic pricing, they chose such consumer goods as automobiles and refrigerators. For example, in studying hedonic prices for new autos Griliches turned his attention to changes in such parameters as horsepower, weight, length, and whether they were hardtops or convertibles, had automatic transmissions, power steering, and power brakes. Evaluations were obtained for these parameters on the basis of observed price statistics concerning use of old and new automobiles. Some of the parameters, for example, presence of an automatic transmission as standard equipment, were evaluated only on a binary scale: one if they had the attribute, zero if not. Certain important variables, like fuel consumption, acceleration, durability, and drivability, were not considered in the research because data collected in the proper way were not available. There are data for these variables, but they are scattered among different sources. In principle these data could be collected, and values for the appropriate parameters could be constructed.

Peter Linneman (1980) conducted some highly interesting research into determining hedonic prices for U.S. housing. Based on a study of statistical data for the country as a whole and for several individual regions in particular, the author attempted to discern the influence of different factors on the formation of housing prices. Among these factors were age of the building, number of rooms, number of floors, whether there was an elevator, was the main bedroom a walk-through, was the kitchen fully equipped, had it been recently flooded, number of toilets, was it connected to a sewage network, did it have central heating/air-conditioning, did the building have a trash collection system, did the house have a basement, were there holes in the floor, did the building provide exterminator services, were its hallways illuminated, the number of toilets without showers, was there a garage, were there too many bad smells in the neighborhood, was there too much business activity in the neighborhood, were the streets bad, were the side-streets well lit, was the crime rate too high, were the schools considered bad, were the neighborhood stores considered bad, was there too much traffic noise and airplane noise, was traffic too heavy, were the houses too crowded together, and were there too many trucks or too much garbage in the neighborhood. Thus, all the conditions for forming hedonic prices began from the fundamental position that it is possible to take into account, in the price of an individual good, both variations concerning its internal structure and complementary products external to it.

Setting hedonic prices parallels the procedure for computing predispositions if only because in both cases the base price of the product is adjusted based on certain value-assigned parameters. However, the two methods differ profoundly. Computation of predispositions proceeds by taking a set of products and their unconditional prices and supplementing it with a set of relations among the products. Moreover, it is assumed that a single procedure cannot interlink all the relations arising during the course of the system's operation in a consistent and complete manner. Hedonic prices are set by taking as a base price the conditional price of an existing good, that is, a price determined by all the relevant circumstances. This price is subsequently modified based on the value of parameters incorporated into the new product and, of course, lacking in the old one. In other words, hedonic prices presuppose that the price of the base product does take into account all the various relations arising during the course of marketing

production-acquisition of the goods but that this price must be adjusted to reflect features peculiar to the new product.

By no means do I wish to undermine the significance of hedonic prices. I want only to underscore that these prices are a variety of ordinary equilibrium prices under determinism and that their application makes sense wherever equilibrium prices are applicable.

ESSENCE OF THE AESTHETIC METHOD

Now, with all the different characteristics of the potentiating method laid bare, I would like to reexamine the basic essence of this method as it pertains to aesthetics. Let us look first at a number of examples dealing with the evaluation of results in the various fields of human endeavor.

Our journey begins from afar — from chess. The development of chess theory led to the appearance of the positional style toward the end of the nineteenth century. Before that the combinatorial style was predominant, characterized by finding a limited, clearly formulated goal, usually the gain of material, and then elaborating a program for its accomplishment. The accomplishment of this goal practically guaranteed a win in the whole game. The positional style was characterized, above all, by the formation of the best possible position under circumstances where one cannot tell specifically how the game will proceed. The positional style does not abolish the combinatorial one; it merely anticipates it.

The computerization of chess has substantially advanced the game's positional style in the sense of formalizing it. The honors doubtlessly belong to Claude Shannon. According to Shannon (1950), the weight function f for some chess position P could look as follows:

$$f(P) = 200(K - K') + 9(Q - Q') + 5(R - R') + 3(B - B' + N - N') + 1(P - P') - .5(D - D' + S - S' + I - I') + .1(M - M') + \ldots,$$

where

K, Q, R, B, N, P are White's extant king, queen, bishops, knights, and pawns;
D, S, I = doubled, backward, and isolated white pawns; and
M = White's mobility (measured, for example, by the number of legal moves available to white pieces).
The said letters followed by an apostrophe (K') refer to black pieces.
The coefficients 200, 9, 5, 3, and 1 are the widely accepted valuations of relevant chess pieces; the coefficients .5 and .1 are the rough-and-ready valuations of positional parameters proposed by Shannon.

Shannon's formula can be represented as a linear polynomial function

$$M = C + O$$

where

C = the product of the quantitative differences of the material parameters K, Q, R, B, N, P of each player and the evaluations of these pieces, and

O = the product of the quantitative differences in the positional parameters D, S, I, M of each player and the evaluations of these parameters.

Of enormous interest from the general methodological perspective is the fact that chess is, apparently, precisely the only area where there has been any success in vividly revealing in a rigorous fashion the necessity of unconditional evaluations for material parameters and to construct them analytically.

Chess players distinguish two types of valuations: conditional and unconditional. Conditional, or equivalently situation-specific valuations, incorporate all the constituent elements of a given position relative to the set goal. In other words, these valuations denote the contribution of a given piece to the attainment of the set goal under certain constraints. The relative valuations of the pieces widely accepted by the chess players (queen = 9 pawns, rook = 5 pawns, bishop and knight each = 3.5 pawns) are precisely the unconditional valuations mentioned above. In turn, these valuations formed in an idealized setting, which reflects the ability of each piece to ensure one necessary condition required to win the game (Taylor, 1876).

The following example should help an uninitiated reader understand the difference between conditional and unconditional prices. It concerns a chess problem usually given to beginners. Suppose we have a position on the board (Figure 8.1) and it is White's turn to move. The player with the white pieces knows the rule that allows the pawn to be promoted to any piece upon reaching the eight rank. The beginner's first impulse is to promote his pawn to a queen, for he knows the queen is generally the strongest piece. But if he falls victim to his natural impulse, black will mate white in one move with his own queen (Figure 8.2). However, if White were to promote his pawn to a knight, black king would be checked and mated immediately (Figure 8.3). The difference between conditional (situation-specific) and unconditional valuations should now be apparent.

The results of the search of positional parameters and their evaluations reflect to a large extent the creativity of a given player.

FIGURE 8.1
Initial Position

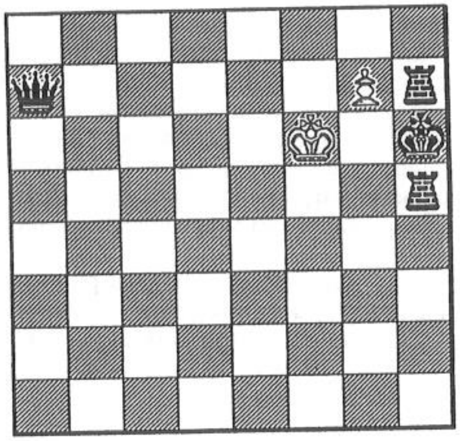
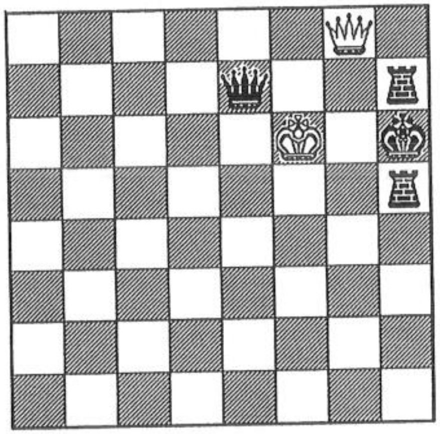


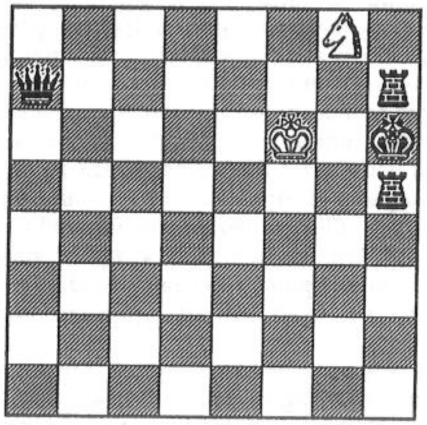
FIGURE 8.2
The Tempting Move



These evaluations are constructed subjectively because who will be realizing the examined position is of extreme importance. This does not exclude their statistical study as an auxiliary means.

The next example has to do with economics, particularly with a firm. This could equally be applied to a branch of industry or the economy in general. Jumping ahead of myself, the above discussion

FIGURE 8.3
The Winning Move



of the potentiating method suggests the following. According to classical theory, the estimate of the behavior of a firm is based on profit. The notion of the impossibility of completely tying in the present state of the firm with its future supposes that there are no complete (conditional) prices on the basis of which profit could be measured at any given point in the remote future. In this case a more complex procedure is needed to evaluate the state of the firm. First of all, the value of the output and input is calculated in unconditional prices, and then these values are subtracted from each other. Then one takes into consideration as independent variables the parameters of relations with the corresponding evaluations. These parameters are equally relevant to the output of a firm (for example, the firm's share of the market or the completeness of the manufactured set of products) and the resources consumed by the firm (for example, the correlation of vertical and horizontal integration in the firm). The evaluations of the parameters of relations can be either positive or negative.

Hence the formula for evaluating the firm's predisposition to development is:

$$M = C + O$$

where

M = the measure of the firm's sytropy,
C = the difference between the values of input and output,
O = the value of the parameters of relations.

In its turn,

$$C = I - E$$

where

I = benefits from material parameters and
E = costs of material parameters.

In its turn,

$$I = SP_{\mathrm{M}}$$

where

S = the vector of outputs,
P_{M} = the vector of unconditional prices on products/resources, and
E = IP_{M}

where

I = vector of inputs and
P_{M} = the vector of unconditional prices on products/resources.

$$O = RP_{\mathrm{R}}$$

where

R = the vector of relational parameters and
P_{R} = the vector of valuations of relational parameters (with positive and negative components).

It is immediately noticeable that the independent variables of the weight function of a firm's position are the same material (in this case essential) parameters and positional parameters (in this case parameters of relations) as we previously observed in the evaluation of chess. This sort of isomorphism in the evaluation of a firm's position and a chess position should hardly be surprising. In both situations we are dealing with what is in some sense a game situation under conditions of breaks between its initial state and its future (or past) development.

Things are less familiar with the weight function measuring the beauty of a work of art. The honor of constructing it belongs to the well-known U.S. mathematician George Birkhoff. His fundamental work in this area was published in 1933. In subsequent years this book provoked great interest, expressed not merely through corresponding theoretical studies but also through attempts at an experimental verification of its theses (Smets, 1973).

Birkhoff's formula for the aesthetic measure M is a function of two variables, C = complexity and O = order.

Here is how he clarifies these measures:

> The typical aesthetic experience may be regarded as compounded of three successive phases: (1) a preliminary effort of attention, which is necessary for the act of perception, and which increases in proportion to what we shall call the complexity (C) of the object; (2) the feeling of value or aesthetic measure (M), which rewards this effort; and finally (3) a realization that the object is characterized by a certain harmony, symmetry, or order (O), more or less concealed, which seems necessary to the aesthetic effect (pp. 3–4).

To illustrate his ideas Birkhoff uses the example of a convex polygonal tile. The measure of its complexity is determined by the number of its sides and the measure of order — by such parameters as repetition, similarity, contrast, equality, symmetry, balance, and sequence.

Formally, Birkhoff's variables look as follows:

$$C = ra + sb + tc + \ldots,$$

where

a, b, c = indices of tension of adjustments A, B, C (that take place r, s, t times) that the nervous system must perform for the essential parameters to be perceived. These indices have a negative sign because perception here is impossible without sustained interest.

$$O = ul + vm + wn + \ldots,$$

where

l, m, n = indices of tone of feelings that correspond to different types of associations L, M, N repeated u, v, w times. Feelings could be positive, negative, or indifferent. Ambiguity, undue repetition, and unnecessary imperfection produce clearly negative feelings.

The aesthetic measure M is a ratio of the above two variables O and C:

$$M = O/C$$

Although I have high esteem for Birkhoff's approach to the construction of a criterion of aesthetic measurement, I would like to offer some comments on this criterion, beginning with the general appearance of Birkhoff's formula for aesthetic measurements. Some authors disagreed with this formula and held that the expression $M = OvC$ is more appropriate for the measurement of a degree of aesthetic value (Eysenck & Castle, 1970). These critical considerations were primarily founded on disagreement with Birkhoff's implicit notion that simplicity and symmetry define human perception. If one were to follow the form of criterion suggested by Birkhoff, it might turn out that a more complex and more organized object will possess the same degree of aestheticism as a more simple and correspondingly equally well organized one.

According to Birkhoff, all the complexity parameters have a negative valuation whereas order parameters can be either positive or negative. In principle, the more aesthetically evolved an object, the less it will possess negative parameters of complexity. I favor the view that both kinds of parameters, to be perceived as information signals, require an expenditure of nervous energy and that these expenditures could have a negative sign. Having passed through our nervous system, however, these signals agitate the nervous centers that could produce both positive and negative emotions.

These last remarks will be better understood after we have looked at the analogy that Birkhoff draws between his aesthetic formula and economics, saying that a business's success is determined by the ratio of investments to profits. In itself the latter statement is true, but some clarifying comments are needed to make the analogy more precise. The opposition of investment to profit in economics is dictated by the presence of resources with short-term and long-term service life. The acquisition of machinery requires a single expenditure, but this machinery can work for a long time. Similarly, a protracted production cycle provides working capital. In economics the opposition of investment to profit can be eliminated if we include the interest rate in current costs and use them to calculate both cost and profit.

Birkhoff's analyzed analogy between evaluating the success of an enterprise's activity and aesthetic degree can be deepened if we take into account the above cited comments on evaluating a firm's predisposition for development. There seems to be a complete analogy here: complexity is an analogue of value of material parameters whereas order is value of relational parameters.

But the issue is not in the outward analogy and isomorphousness of mathematical formalisms that join the weight functions in economics, chess, and aesthetics. The homogeneity of a mathematical formalism serves only as a pretext for pronouncing a hypothesis on the homogeneity of phenomena concealed behind them.[8]

Below I will attempt to demonstrate why calculating predispositions can be essentially considered aesthetic measurement aimed at clarification of the object's degree of beauty. Naturally, this does not exclude the particularities of calculating the beauty of artworks, in nature, and in systems artificially created by people. Thus, calculating the beauty of artworks according to Birkhoff is based on immediate clarification of the level of excitation of our nervous system, that is, the excitation of those structures that apparently store the information of value of the elements forming images or perhaps of the simplest Gestalts.[9] The difficulty of aesthetic measurement of artifacts was apparently related, above all, to the requirement of constructing its own value system demanding new axiological principles. This is why, in my estimation, aesthetic measurements are so scantily developed in many areas of human activity where it is impossible to access directly fundamentally profound values.

Before examining the essence of the aesthetic method, I would like to make two propositions. The first proposition concerns the scope of the aesthetic method. Traditionally, this method has been applied to artworks, and it is with regard to them that it has been most fully developed. Still, this does not preclude its application to other objects, both natural and manmade. This wider application, however, has been quite unsystematic and desultory. Thus, a mathematician can talk about the beauty of a theorem's proof and a physicist about the beauty of a constructed physical theory but in doing so they will characterize this beauty sharply and, moreover, in terms that, in turn, need to be defined, for example, elegance. The beauty of nature or human behavior is characterized by much the same approach, hence, my first proposition: the aesthetic method, under appropriate conditions, can be just as effective in analyzing any object in any field as it is in dealing with artworks.

The second proposition concerns the subject who uses the aesthetic method. Commonly, it is thought to be a prerogative of humans and, sometimes, of highly intelligent animals insofar as they are capable of producing art. That's why scholars have tried to locate the stable psychological and physiological bases of this method. Individual human valuations, of course, have not thereby been ruled out.[10]

I think that any interacting objects capable of aesthetic perception (and any object of the system so endowed can act as a subject), in any field, can use this method while by-passing human involvement.

A human still has to be there, of course, to analyze what's going on. But there is a big difference between a person's direct aesthetic interaction with an object and the object's aesthetic interactions between themselves. Admittedly, the aesthetic method is present in interactions between developing cells, in the formation of seemingly counterproductive intermediate genotypical and phenotypical structures, in the interactions of some particles.

Thus, my second proposition suggests that although people produce aesthetic measurements as well as all other sorts of measurements, along with the given gnoseology there is also an ontological aspect; that is, the aesthetic method is used in various systems in the sense that, in the process of the functioning of these systems, the interacting objects themselves produce the measurements of an aesthetic nature.

To sum up, I see the aesthetic method as a general system's phenomenon that can, under appropriate conditions, be applied to any interacting objects in any field. In order to help the reader appraise the importance of this sort of new outlook on the aesthetic method in economics, I offer statements concerning this problem by one of the patriarchs in this area, Russell Ackoff.

As Ackoff (1981) notes, the role of the aesthetic method in the economy is quite great. Following in the footsteps of ancient Greek philosophers, Ackoff includes the creation of beauty as an aesthetic function of society among the four conditions — along with truth, plenty, and good — necessary and, in tandem, sufficient for a person's advancement on the path to omnipotence.

> This function requires more extended treatment because it is the least understood and, I believe, currently the most critical of the four. The mystery associated with aesthetics is reflected in the fact that throughout history very few systemic philosophers have been able to incorporate it in their philosophies. Of the few who have, the results have been more like appendices to their philosophies than integral parts of them. On the other hand, very few of those who have been preoccupied with aesthetics have made a significant contribution to our understanding of science, economics, ethics, or morality. Historically, aesthetics has been the black sheep of the philosophical family.
>
> This disconnectedness of aesthetics is reflected in the fact that managers have some idea of what the science, economics, and the ethics or morality of management mean, but they often have no idea what the aesthetics of management means. It has long been assumed that aesthetics and management have little to do with each other (p. 39).

Ackoff sees beauty as people's ability to stimulate new desires and spur on the striving to seek their satisfaction. Continuing to develop his views on the subject, Ackoff connects them to the problem of goal and means, both in the sense of formation of infinite goals, which always leaves open to people the path of search, and in the sense of value of means independent from the goal.

In connection with this, Ackoff cites certain examples illustrating the role of the aesthetic method in the management of corporations.

> The importance of style in corporate life is obvious. For example, three executives who wanted to diversify the company they owned in order to be challenged more and to become more involved in their business which pretty well ran itself. In their words, they wanted "to get more fun out of it." Fun is a matter of aesthetics and recreation, not science, economics, ethics, or morality.
>
> Another example: a major corporation's profits have been suffering recently because of its commitment to producing only the highest quality products in the field. Its cost of materials is now inflating more rapidly than are its competitors', but it refuses to adulterate its products or abbreviate the processes by which they are made.
>
> Degrading its products would significantly reduce the satisfaction that the managers derive from their work and their pride in the company. This too is a matter of aesthetics.
>
> Finally, there are two separate districts of the Federal Reserve banking system that have exactly the same functions. Nevertheless, they differ significantly in the way they are organized and carry out their functions. The atmospheres in these two banks are very different. The differences cannot be explained in scientific, economic, or ethical-moral terms; they are aesthetic, or stylistic, in character (p. 42).

Ackoff sees enormous difficulties concerning the aesthetic measurements of the quality of life by analytical means. Therefore, he suggests realizing this measurement through the participation of people in the management process, thus tying in his approach to management with the idea of participatory planning.

Thus, on the whole Ackoff's attitude toward the aesthetic method remains within the framework of the classical paradigm, which ties the aesthetic method to the perception of beauty by people through their emotional mechanism. But even within the framework of such an approach, the value of which is indubitable, it is missing an extremely important link dealing with the enrichment of a worker's intuition by analytical vision of the beauty of the system the worker manages. This vision does not deny a final evaluation by people of their own activity through the emotional mechanism; on the contrary, it can greatly enhance it.

Our discussion of the potentiating method and its relation with aesthetics can be extended to the socioeconomic domain, that is, the interplay of economic and social phenomena. Because relations are organically incorporated into the potentiating method, it would make sense to evaluate the state of an economic system based not only upon the material parameters intrinsic to the economy itself but also upon the relational parameters, especially the human relations. What distinguishes the latter is that they are personal, unlike the subject matter of pure economic science, which studies impersonal relations, in the sense of disregarding the recipient of a particular good or service.

Investment, such as education or health care, in an individual has given rise to a distinct and practically useful branch of economic science called "the investment in human capital." The field of investment in human relations, however, is still a prerogative of a few scholars involved in the study of a rather broad concept known as grants economics.

The concept was founded by Kenneth Boulding (1981). Grants economics basically involves one-way transactions of goods and services, as opposed to traditional economic science concerned with two-way transactions, that is, exchange of goods and services collectively termed "exchangeables." Grants economics encompasses diverse areas ranging from taxes to an invitation to a private dinner party. Consequently, grants economics could well include investment in human relations, a peculiar subject calling for an in-depth study.

Regrettably, this concept has not been widely acknowledged in the more practical circles. Thus, many conflicts (often concealed) occur in corporations among people at different levels of the managerial hierarchy.[11]

NOTES

1. The term object can be applied to prices as well to underscore the fact that prices can vary independently just like primary objects. In other words, prices are not shadows, as they are often referred to in mathematical economics, but objects in their own right that can be manipulated directly thus affecting the behavior of primary parameters. A deeper analysis of the problem of value is presented in my book (1984).
2. These lists of parameters were developed in corroboration with my former students Joel Friedman, John Lisowski, Kent Myers, and Jacob Schultz.
3. The prevailing classification of the valuations into intrinsic and extrinsic or instrumental is largely characteristic of the described parameter grouping in the conditions of determinism and indeterminism. Intrinsic values

pertain to indeterministic situations where insurmountable gaps in the network stipulate the need to measure the potential of a given state. Unlike intrinsic values, extrinsic (instrumental) ones presuppose complete and consistent links between different states of the system. It seems to me that it is common to interpret intrinsic values in the extreme sense, meaning the valuation of the object as such — its immanent quintessence. The same holds true for extrinsic value understood as being completely predicated by the specific objectives.

4. All these considerations have led me to conclude that the quantum mechanical representation of a system, which in physics is associated with the uncertainty principle and indeterminism, is just one particular case of a general indeterministic representation of a system. Indeed, quantum mechanics assumes that, in principle, all the necessary information to determine a particle's trajectory exists. However, this projection is fundamentally impossible to realize because the primary and conjugate parameters (the particle's position and its impulse) cannot be simultaneously determined.

In order to find the trajectory of a particle, it was necessary to revise the classical laws. Schrodinger's equation accomplished just that. Let me venture into very speculative territory and hypothesize that, perhaps, the gaps in this equation between the information on the particle's position and the particle's impulse are bridged primarily by the valuations of the relations among the particle's represented in a very condensed form as Plank's constant. The impetus to my speculation is the fact that as we approach the speeds of classical mechanics, when the gaps between measuring the position and the impulse of an object can be ignored, Plank's constant becomes irrelevant. Perhaps, the solution of Schrodinger's equation reflects the predisposition of a particle to appear at a certain position, rather than the widely held interpretation of this value as a measure of probability that a particle will appear at a certain position. Using statistical information to evaluate the possibility that a particle will appear at a certain spot really implies a search for the probability of its appearance. If we could reasonably assume that, under certain conditions, predisposition converges to probability, it would add credibility to the possible agreement between the results of calculating the probability that a particle will assume a certain position and the results of measuring its predisposition to assume the same position.

It seems to me, the quantum mechanics situation is drastically different from the more general case of indeterministic behavior of objects. The latter is predicated on the assumption that the pair of parameters that characterize an object reflect incomplete knowledge of the conjugate parameters (complete valuations). I believe the elaborated method of potentiating is special, in that it allows us to find the degree of indeterminism in this more general case.

As long as I am head over heels in speculation, I will take a chance and say a few words about the connection between quantum mechanics and gravitation, the general field theory. Perhaps, in this realm, the missing links in the classical model call for a more sophisticated method of filling the gaps than Plank's constant. Possibly, the super string concept which is based on multi-dimensionality of space — the introduction, together with four-dimensional space, of six additional dimensions — is an expression of the relational parameters I have mentioned as independent variables. But if these additional dimensions in physics are a consequence of deduction in group transformation, then in economics (as in chess) they are perhaps evidence of an as yet unidentified group transformation.

5. Confirming the conservative nature of human relations is our present-day perception of Greek drama and Shakespearean tragedies. To this day they seem so relevant that some producers stage them using contemporary costumes.

6. A very far-fetched speculation, but perhaps a linear polynomial, which denotes predisposition, might help in solving certain nonlinear problems irreducible to linear cases.

7. The simultaneity requirement is often only partially fulfilled, that is implemented by means of independent procedures dictated by particular concerns. For instance, aggregation of different products and their respective prices is vital in computing the price index. The process of disaggregation comes to the fore in setting prices of new kinds of goods.

8. Just that sort of phenomenon took place in the nineteenth century when James Maxwell, on the basis of the discovered homogeneity of differential equations describing electrical, magnetic, and light waves, demonstrated their common nature.

9. The evolution of Wassily Kandinsky (1866–1941), founder of abstract painting, reveals the artist's striving to discover certain aggregation in previously discovered basic elements. It was expressed in Gestalts possessing the simplest biological forms. It was precisely the search for these that characterized the last, Parisian, period of the great painter's life.

10. The latter statement acquired particular significance after the discovery of emotions as special physiological structures of the brain. Earlier, emotions were considered only a psychological phenomenon, a product of the activity of the brain's other physiological structures. Emotions should be distinguished from evaluations on the basis of which an individual ultimately arrives at decisions. In forming the latter, many other conditions are taken into account.

11. The following story comes to mind in this connection. I taught a course in a program called the Dynamics of Organization at the University of Pennsylvania. At the beginning of the course, I required all my students to write a paper expounding their ideas regarding the way they made decisions and the way it meshed with the decision-making mechanism for the organization as a whole. If, by the end of the course, a student had gained some new insights, the paper had to be rewritten; if a student could not do so, the fault of not providing any new knowledge lay with me.

One of my students worked as a manager of a regional marketing branch of a firm that assembled trucks from parts supplied by other specialized companies. In his paper, this student made a vehement assault against the decision-making policy in his company. He wrote with indignation that rather than plan the company's future, company executives spend a great deal of time going to lunches and dinners, just having a good time. I expected the final note of the paper to read that the company was either on the verge of bankruptcy or already in bankruptcy. Instead, I read that the company's profits were rising from one year to the next.

What was the reason for the inconsistency in the picture painted by the student? As I have mentioned already, this was an assembly company. The two crucial factors responsible for its successful operation were getting more orders for its trucks and having reliable suppliers of component parts. The assembly process itself was not a constraint because the firm was cushioned by the 120 days it required to fill an order. Under these circumstances, company executives deemed that establishing personal relations with clients and suppliers of parts was their primary task, which they fulfilled during those lunches and dinners.

9 Macroeconomy I: Measuring Economic Potential

DEFINING THE POTENTIAL OF AN ECONOMIC SYSTEM

Strictly speaking, potential is most relevant when we speak of development. However, it also has limited application to growth and survival because the latter, at least to a degree, represents a particular case of development. Potential is a characteristic of the state of a system.

In the broadest sense of the word, this concept alludes to hidden possibilities. Webster's dictionary defines potential as something "existing in possibility." Dissecting the structure of the concept of the potential with "possibility" understood in the broadest sense of the word, it becomes apparent that potential incorporates not only the system's ability to transform one object(s) into another but also the forces that evoke and guide these efforts.

With this in mind, our approach to the category of economic potential should incorporate both the desires of the people to produce goods and services as well as the means at their disposal to do so. Lack of either one of these two conditions strips the system of its potential.

Let me elaborate upon what I mean by desire. In general, economics deals with two kinds of product valuations. One reflects an equilibrium (optimal) state and, in terms of an equilibrium model, assures us that demand and supply are in equilibrium; in the lingo of the optimization approach, it means that, given goal and constraints, we have maximized the value of the criterion. This variety of valuations may be viewed as being conditional. The other kind of valuation is a direct reflection of the utility function. It shows the utility of a particular product to the individual independent of the amount of the good that is currently available. These evaluations can be thought of as being unconditional. If the utility function is convex, a necessary condition for growth and development is for the current

valuation of a product to exceed its maximal possible unconditional valuation, which in this case equals zero. In order for growth and development not to stop — a case where unconditional and conditional valuations become equal — we must constantly search for new desires that can be fulfilled. The problem is further simplified because a segment of our utility function is nonconvex (such as craving for power or knowledge); in this case, the unconditional value may have a positive correlation with the magnitude of the respective parameter so that it can well be insatiable.

To summarize what has been said, desire is the difference between the valuation of a product under full saturation (in case the utility function is convex) or under unlimited saturation (in case the utility function is nonconvex) and its valuation in the present. In other words, desires represent the driving forces.

Ability encompasses all available resources and methods of transforming these resources into products, as well as the ability to reconcile our desires with all the various traits incorporated into our abilities, and the realization of these desires through actions.[1]

The concept of potential should not be reduced to any one of the aforementioned components; but these constituent characteristics should be viewed as an entity. This kind of approach toward the concept of potential is equally valid for the analysis of global economic development as well as its growth and survival (that is, it applies to the potential for development, growth, and survival). Potential for survival refers to the ability of the system to operate under stationary conditions with production level above zero and a coefficient of growth equal to one. Potential for growth indicates the ability of a system to operate with a growth coefficient greater than zero. Potential for development reflects a system's ability to create conditions that will guarantee an increase in its growth rates in the future.

Having described the multifaceted nature of survival, growth, and development, we are in a position to define the concept of the potential for development in more concrete terms using the systems approach. Let me reiterate the point made in Part I.

From the functional point of view, potential of the economic system serves to assimilate unforeseen events to the system's own advantage; to prevent undesirable consequences of unforeseen events; and to induce the environment to the system's own advantage.

From the structural point of view, potential for development assumes the presence of material parameters as well as their relations with the latter also being treated as independent parameters. Both the material parameters and their relations must be evaluated appropriately (for instance, material parameters evaluated in unconditional prices).

From the point of view of the process, potential for development refers to the multilevel hierarchy of control methods, control methods for changing control methods and so on. We introduced a multilevel hierarchy as a requirement because control methods available at each level are, generally speaking, incomplete and inconsistent.

The process of potentiating is actually limited to development. For both growth and survival it is reduced to a program that simply links inputs and outputs of the economic system.

Indeed, consider potential for growth. It does not have to be fractured into all the constituent parts characteristic of growth potential because the respective program controlling the operation of the system will have interlinked all the components of the system.

In terms of function, potential for growth assumes that all parts in the system are interlinked and, therefore, the system is fully prepared to face the future. Hence, it is unnecessary to isolate individual functions arising in the course of operation. From the structural point of view, all that is required for growth potential is the presence of material parameters. Their interrelations are not regarded as independent variables because the relations will be implicitly incorporated into the program, which interlinks different parts of the system. Prices of material parameters generated by this program represent Lagrange's multipliers. They provide, in a very condensed form, information regarding all parts of a given system. From the process-oriented point of view, potential for growth is limited to a one-level control program (because this program is complete and consistent) responsible for bringing all the components of the system into an equilibrium.

Potential for survival reduces to a minimum level of stationary growth.

Meanwhile the current literature on economic potential discusses the concept of economic potential solely in reference to the potential for economic growth, not for economic development. Most definitions isolate a single aspect of economic potential — either the driving force or, more frequently, the economy's capacity for output. The prevailing concept of economic potential revolves around economy's output capacity. Some definitions from this class emphasize the economy's capacity to increase output under full (or greater) utilization of available resources. In other words, proponents of this approach feel that because some resources available to society remain idle, we need a new category that would reflect the output frontier under full utilization of resources.

I quote from Paul Samuelson (1966), who adheres to this concept of economic potential:

> *Our Economic Potential.* Had our economy progressed since 1956 — not at the dramatic sprint of the Western European or Japanese economies or at the rust of the controlled totalitarian systems but simply at the modest pace made possible by our labor force and productivity trends — we could have expected 1961 to bring Gross National Product some 10 per cent above the $500 billion level we are now experiencing. With unemployment below 4 per cent, with overcapacity put to work, and with productivity unleashed by economic opportunity, such a level of activity would mean higher private consumption, higher corporate profits, higher capital formation for the future, and higher resources for much-needed public programs. Instead of our having now to debate about the size of the Budget Deficit to be associated with recession, such an outcome would have produced tax revenues under our present tax structure sufficient to lead to a surplus of around ten billion dollars; and the authorities might be facing the not unpleasant task of deciding how to deal with such a surplus (pp. 25–26).

Here, the dynamics of the potential can be measured as the proportions between the yearly percentages of utilization of resources.

A more general approach to this problem was taken by Abram Bergson (1961).[2] Using the Soviet case as an example, he studied economic growth from two angles. One was the aggregate product expressed in terms of national income or as a value of the welfare function (it is immaterial whether this function reflects the interest of the people or of the leaders). Bergson defined the value of the aggregate product as expenditures on resources utilized. He assumed, based on general theoretical considerations, that prices of resources reflect their productivity. In connection with this last aspect of input productivity, Bergson introduced the notion of production potential: "Production potential at any time is understood in terms of the corresponding 'schedule of production possibilities,' that is, the schedule representing the limiting amounts of 'composite commodities' of different mixes that the community might produce with available factor supplies and 'technical knowledge'" (p. 26).

On the basis of available Soviet data, Bergson carried out a thorough analysis of the various ways of measuring production potential. Using his broad theoretical background, Bergson precedes his empirical findings with an extensive examination of general methodological principles involved in computing constant prices. Here, Bergson could not help falling into a very common trap that poses danger to anyone who attempts to measure economic dynamics using constant prices. With everything that has been said regarding the measurement of economic growth and development, we see that Bergson really attempted to sidestep the problems of measuring economic growth. Conceptually, he remained within the orthodox framework unable to free himself from its limitations and examine the problem in a more general context of economic development.

To sum up our discussion of economic potential as some feasible space of production, the use of the term potential in this context seems to me inappropriate — even when it is accompanied by such epithets as economic or production. Specific terms in economic literature reflect this aspect of the economy — production possibilities frontier and feasible production space.

It is not surprising that a rather unusual and original definition of economic potential was proposed by Kenneth Boulding (1970). By analogy with physics Boulding's approach incorporates the notion of driving forces. The following passage illustrates his basic idea:

> Wherever we find a potential difference producing a current or flow by overcoming a resistance, we find something like Ohm's law, exhibited in purest form in the study of electricity, but valuable as an interpretive principle when we study the flow of goods or resources in response to price differences (economic potential) against the resistance imposed by costs and transport. In the theory of electrical circuits we may find clues to some baffling phenomena connected with the circuit flow of money (p. 8).

Boulding's analogy concerning potential and the value of goods at their production and consumption sites, respectively, was widely used in economic literature dealing with linear programming and transportation problems in particular. In solving these problems modeled after a hydraulic system, Leonid Kantorovich (1965) used the concept of the potential to describe fluid levels in the system of communicating vessels. Boulding's analogy between economic potential and electricity was used directly in designing analogue computers as electric networks adapted to solving certain classes of problems in mathematical programming (Dennis, 1959).

Boulding's ideas regarding economic potential are most applicable in an equilibrium approach, which corresponds to a stationary regime of operation or, in other words, to the potential for survival. Other conditions being equal, deviations in the difference of the potential charges (in the physical sense of the word) might lead to the system's collapse. Boulding's definition of economic potential makes sense in growth measurements only under complete aggregation of all the products comprising the economic system. As Boulding points out, aggregation can be achieved by representing all products in monetary form. Boulding does not indicate how to measure the growth of the potential of an economic system comprised of many diverse objects (unlike a physical system having a uniform current flowing through a single channel). All problems regarding the measurement of the potential for development are completely foreign to this approach.

MEASURING ECONOMIC GROWTH

Social Aspects of Growth Measurement

One very strong assumption is incorporated into measuring economic growth: all interrelations among the various parts comprising an economic system, from the development of novel ideas to the production of consumer goods, are known. Even with this condition fulfilled, a number of problems arise in measuring growth because social implications of economic activity have a feedback on economic development.

If nature's gifts were gathered for consumption as final goods by independent groups of people, measuring economic growth of such a system would not pose any problem; growth would consist of the total value of the final goods gathered. Nevertheless, even this simple case gives rise to a number of social problems related to the choice of the composition of the final basket of goods.

Economic systems consisting of elaborate production networks present many more problems. Output at each moment of time should include not only final but also capital goods whose contribution to the output of final goods is counted only at some future point. The category of national income reflects precisely this vital aspect of computing economic growth.

This aspect of growth measurement results in a number of socioeconomic problems, for it allows the production of the final goods to grow nonmonotonically. It may actually be beneficial to forgo consumption of the final goods and utilize the freed resources for the production of capital goods. This process may displease people who would rather consume today than save for tomorrow.

What complicates matters even more is the development of science and its applications. Suppose we are in a position to measure the impact of science on technology and, consequently, on the value of the criterion of optimality. Under these circumstances restructuring the existing production network may actually lead to the curtailment of production, not only of consumer goods but also of some previously produced capital goods. For instance, it may be expedient to cut production of certain goods and channel the resources into other products. This restructuring may give rise to unemployment: laid off workers may have difficulty gaining employment elsewhere either because they lack the necessary skills or simply because of insufficient funds to occupy all of them at new jobs (the problem is aggravated for the workers' families). Thus, restructuring the economy may allow some workers to resume their work, but others may be forced to become more skilled or change their occupation altogether.

Implications of measuring economic growth and development are socially very important. Misjudging the social component (especially underestimating it) resulting from structural-technological shifts in the national income may lead to an even greater sociopolitical unrest, which accompanies development.

The Microeconomic Optimization Approach to Measuring Growth

There are basically two methods of measuring the national income of the economic system as a whole: macro and micro. A macroeconomic approach based on a hierarchically organized set of aggregate parameters is an indispensable tool in managing an economic system. It is employed primarily to solve general problems having a relatively small number of parameters (essentially this characterizes any large-scale system). At the same time macrodecisions must be implemented at the microlevel. Therefore, the process of parameter aggregation carried out at a macrolevel must be accompanied by disaggregation at a microlevel. Eventually, the latter step serves to implement factor relations generated at the macrolevel at the microlevel, which is characterized by local problems having a small number of detailed parameters.

I emphasize eventually because this step represents an interactive disaggregation procedure in the course of which the values of microeconomic parameters may change, which, in turn, affects the process of aggregation. The reverse takes place during aggregation, which affects the values of macro parameters. This highly interesting problem of designing iterative aggregation and disaggregation procedures is discussed in several publications (Dudkin, Rabinovich, & Vakhutinsky, 1987).

I focus on the problems of measuring economic growth (as well as survival and development) arising specifically as a result of using a microeconomic approach to analyze an economic system as a whole.

Our investigation depends on the way we represent a microeconomic system. Chapter 5 introduced two classes of methods — reactive and selective. The first class represents an economic system analytically. Statistical analysis of the economy over time also gives rise to this kind of representation. A few problems are associated with this approach, first and foremost, finding the function that brings the analytical representation. The value of this function represents an output vector of intermediate and final products and its arguments are the inputs — labor, natural resources and intermediate goods. Furthermore, growth in this case can be measured only by comparing the values of these vectors, which, in turn, requires that we assign weights (prices) to its components. But, because prices are also based on statistical data, they must be stripped of the effects of various "flations" — inflations and deflations. This creates a hopelessly entangled maze. In order to eliminate the effects of flations, it is necessary to introduce a standard basket of goods, in other words, to fix a standard set of products taken in certain proportions. But how do we form this basket? Which year should we take as the base point of our calculations? We know that consumers' tastes as well as prices vary from year to year. New products are introduced while others are taken off the line. These questions are left unresolved.

There are two kinds of selective methods. One uses equilibrium representation, and the other uses optimization. For my purposes, I have adopted an optimization approach to represent an economic system. It is better suited for answering questions concerning economic growth and development. Equilibrium representation does not exhibit growth explicitly, so it must first be converted into an optimization model.[3]

Having said this, let us proceed to examine the problem of measuring economic growth and development within the framework of a microapproach based on optimization. Defining growth would be simple; in fact it would not be any different from development — if our economic system consisted of one single product, that is, if growth could be measured in real units. But the general case of many products poses considerable problems.

Measuring Growth of a Closed Economy

John von Neumann's work (1945) on models of dynamic equilibrium solved this problem as far as multiproduct closed deterministic economies having no technological progress are concerned. Under these assumptions, the maximum growth rate of production is constant, and its sole determinant is the productivity of a fixed set of technologies. Here, growth rate for the system as a whole is the same as that of its individual parts. Von Neumann showed that the growth of a system operating under dynamic equilibrium can also be measured on the basis of its dual characteristics: rate of growth corresponds to the rate of price deflation.

If the state of a system is represented by one single indicator expressed in monetary units (computed by multiplying quantities of outputs x by their respective prices p and taking the sum), all we have to do to determine the system's rate of growth is to fix some arbitrary time interval $(t, t + 1)$, find the volume of output of all the products at time (t) and at time $(t + 1)$, multiply them by the same price vector corresponding to time (*t*), and calculate the ratio between the value of output at time (*t*) and (*t* + 1).

$$p_t x_{t+1} / p_t x_t = C$$

A closed economy does not preclude development, for it could have second-order technologies for changing the technologies. In one particular case when metatechnologies are fixed, their capacity to modify first-order technologies will eventually be exhausted. This will lead to a new constant rate of growth based upon the newly created technologies. Here, an increase in the rate of growth indicates an advance in the potential of the system, in other words, its development. Measuring development in the general case of changing metatechnologies and dynamic equilibrium based upon a constant rate of acceleration of the rate of growth is analogous to measuring growth in a closed model lacking technological progress.

Measuring Growth of an Open Economy

Measuring economic growth of an open system is a more difficult task. Assuming that an open economy has no constraints on outside resources being limited solely by the initial constraints on intermediate goods corresponding to the state of dynamic equilibrium, that the proportions between consumption (products leaving the system) and production are fixed, and that technologies remain unchanged, this type of economy may experience a constant rate of growth in real terms as was the case with a closed economy. Because we measure growth of a system by comparing its states (its total output of intermediate and final goods) at different moments of time, finding its rate of growth is relatively simple. First we have to reduce (divide) the productivity coefficient by the coefficient representing the ratio of consumption to production. For instance, if the productivity coefficient equals six and the proportion between consumption and production is three to one, the system's rate of growth is two.

Relax some of the assumptions made above and consider the following case. Let initial conditions, preferences for the various goods on the part of each individual participant, and the integrating principle for the system as a whole be set arbitrarily. Measuring economic growth in this case requires this integrating principle to be represented in the form of a global criterion of optimality extended over an arbitrarily long period of time.

The difference in the values of this criterion taken at any two moments of time serves precisely to measure the growth of the system. Here it is assumed that the value of the criterion accounts for the fact that at each moment of time the system produces not only final but also intermediate goods that are absorbed in the production of final goods only at some later moment of time.

Under an optimization approach, the values of the criterion of optimality at any two moments of time are sufficient to determine the amount of economic growth that has taken place during that period. We call this approach to growth measurement global because it is based upon the value of the global criterion.

In some cases growth can be measured on the basis of local information. This approach by-passes the global criterion of optimality. It consists of measuring the volume of output at two moments of time by utilizing information regarding the state of the system at each respective moment of time.

The concept of optimality tells us that the total output of final and intermediate goods at time (*t*) is equal to the sum of expenditures at time (*t* − 1) if both output and input are measured in optimal prices. In order to measure the increase in the value of intermediate goods for a particular time period, we must measure, using constant prices, the physical state of the system at two adjacent moments of time.

The theory of optimality specifies cases when this type of measurement is plausible. We discussed one particular case of a closed model of dynamic equilibrium. The solution was simplified by the fact that the rate of growth and the rate of price deflation were the same for all goods. In general, of course, this condition does not hold. Still, we know that prices corresponding to an optimal state (they are also prices of equilibrium), being Lagrange's multipliers, are relatively stable under slight variations in the constraints. This property of prices, together with the assumption that the difference between the states of the system at two adjacent moments of time is relatively small in real terms, allows us to evaluate the state of the system at each respective moment of time using optimal prices corresponding to the initial state. Under these conditions the value of the increase between these two moments of time should be close to the increment in the value of the global criterion for that same period

$$p_{t-1}\, x_t - p_{t-1}\, x_{t-1} = \Delta U_{(t-1),(t)}$$

Knowing optimal prices allows us to construct a chain index and use it to measure growth over an arbitrarily long period of time. The above procedure does not apply to time periods exceeding the durational stability of optimal prices.

Technological Progress and Growth Measurement

At the present time the many problems of measuring growth and development stem largely from our inability to determine all the connections among the sectors of the economy. The R&D sector poses special difficulties.

Current techniques of evaluating intermediate and final products that have passed through the market produce fairly good approximations to equilibrium prices; that is, they reflect both costs and benefits. Such products as R&D, either because they represent public goods or because their benefits are hard to assess, are evaluated in terms of cost alone, and this diminishes their overall contribution. A number of economists realize the need to account for the impact of such products as R&D, and some have tried to make applied macrolevel measurements of the rate of return on investment into R&D.

Meanwhile, return on basic R&D remains an unexplored terrain. At the same time, correct evaluation of basic research is essential for comparing the level of development of different countries. Take the United States and Japan, for instance. The United States conducts intensive research into pioneering ideas in basic science. In such research Japan lags far behind. Indirect evidence for this is the number of Nobel Prize winners from the two countries. The United States has dozens of laureates whereas Japan has only few (and all of them have worked at Western, and in particular U.S. universities for many years). At the same time Japan has surpassed the United States in many technologically advanced fields where what was needed was an improvement upon the already existing ideas and their implementation in mass production. Comparing growth rates of the national product for the United States and Japan while counting R&D only as an expenditure creates a distorted picture of the Japanese advantage. Counting the benefits of scientific research can change the view drastically.

Considering all the difficulties associated with analysis of an economic system using microparameters, we easily see the advantages of evaluating it at the macrolevel.

There are two ways to measure output growth of the final products at a macrolevel in aggregate indicators. One approach is based on inductive reasoning: essentially, we try to construct a time series for various aggregate indicators representing inputs and outputs of an economic system. Based on the analytical expression generated by econometric techniques relating different variables comprising the system, we try to estimate the impact of each individual factor included in the inputs. In other words, the inductive approach emphasizes the idea of finding a function (analytical expression) linking economic parameters on the basis of available statistics. An instance of such an expression is the production function relating an aggregate parameter, such as output of final goods, with the inputs of all the factors that go into making them, including R&D (Griliches, 1987).

$$(d \log U) / dt = a + b(d \log X) / dt + r(R/U) + du/dt,$$

where

U = some measure of output,
X = vector of standard factors of production (man-hours, structures and equipment, energy use, etc.),
a = other factors that affect output,
u = random, unsystematic fluctuations in output,
r = $dU/dK = c(U/K)$,
K = measure of accumulated results of scientific research,
R = net investment into K.

Although this approach does not yield to a direct transition to microeconomic indicators, it is significant in national policy decisions, for instance, the government tax policy to encourage investment into R&D as well as direct government spending in this area.

Another approach is deductive, based on constructing an optimization model linking output of final goods with resources (including technologies for changing the technologies). Here the role of statistics is played by our knowledge of the criterion of optimality given over some set of final products and the technologies linking resource expenditures with the volume of output. In a macroeconomic context these models are known as growth models with endogenous technological progress. The problem with this type of optimization model is the nature of the function linking changes in technological parameters with the costs of technological improvements (including basic science).

Nonetheless, these models are highly helpful in uncovering certain dynamic aspects of economic processes. For instance, a model explored by Ludmila Zelikina (1975), with technological progress given endogenously, reveals rather nontrivial features of the turnpikes under the conditions of dynamic equilibrium with a continuous technological progress.

Under certain conditions one approach can be transformed into another, so these two models based on unlike information can supplement each other. For instance, using optimization models, we can try to establish the relationship between the output of final goods and the resources available. However, it is rather difficult to link R&D expenditures with changes in technological parameters. In the latter case, at least an approximation of the impact of scientific research on the technologies can be made using a production function based on accumulated statistics.

MEASURING ECONOMIC DEVELOPMENT

The economic system as a whole is subject to indeterministic development or to development with discontinuities gaps. How, under these conditions, do we go about measuring changes that have taken place?

The principal answer seems to lie in the idea of switching from measuring growth to measuring development of a system with a certain orientation. This does not exclude growth measurement. It only limits it to stages relative to which we can formulate a goal and link the value of the criterion of optimality with the initial conditions. Suppose the criterion guiding the operation of an economic system is maximization of military power. In this case economic growth should be measured primarily in terms of a system's military growth. Naturally, we can set another criterion of economic performance, such as the standard of living, and use it to measure growth. Setting the criterion can be highly interesting in comparative studies of different economies.

Generally speaking, a distinction between growth and development arises precisely as a result of the indeterministic nature of the world. In a deterministic system these two notions practically merge into one because none of the potentials pertaining to the system need to be evaluated. Each current state can be linked with all other states, so its evaluation reduces to determining its direct contribution to an overall growth. An indeterministic case is very different. It requires any given state to be evaluated in terms of its potential contribution to future development. The essential problem in measuring development is precisely in evaluating this kind of potential.

Let us start by examining the various methods used in assessing the values of material parameters. These values cannot be extracted from any previous state because these states did not account for new circumstances that might arise in the future. Extrapolating future prices on the basis of past prices does not work; it may even be disorienting because real future price ratios may change significantly as a result of some new developments.

A very natural question arises, "Is there a distinct state of the economic system toward which the system gravitates and that is invariant with respect to the available resources as well as with respect to a broad set of criteria of development (intentions of the participants)?" To answer this question, we need to introduce a new category of prices called the unconditional prices. These prices are based on some ideal conditions of future development that the system might approach asymptotically, independently of the changes that might take place — both from the point of view of the intentions of the participants as well as the available resources at the starting point.

This leads us to an important methodological question concerning the representation of a system. I have emphasized openness of an economic system. However, unconditional prices presuppose a closed system. Indeed as long as our definition disregards external inputs and outputs, assuming that the economy consumes everything that it produces and that everything that is consumed is sufficient to start a new cycle of production, it pertains to a closed economic system isolated from its environment. By introducing this abstract notion of a closed economic system, we were able to introduce for unconditional prices a new category that is highly beneficial in an indeterministic representation of an economy.

Turnpike prices, explored by Von Neumann, are a manifestation of this kind of price. Corroborating this affinity are the turnpike theorems proving the existence of a turnpike invariant relative to a broad class of boundary conditions, criteria of optimality, and the initial conditions. Note that unconditional prices do yield to formal analysis. Japanese economists have tried to compute the turnpikes on the basis of statistical data using a highly aggregated set of goods.

An assumption of closeness is one particular aspect of a total economic system that is, generally speaking, an open system. In dealing with specific situations, we introduce boundary conditions (initial or final). This is equivalent to opening an economy because boundary conditions are set outside the system.

We can operationalize the notion of opening up an economy in the following way. The first problem concerns the evaluation of natural resources because their production is outside the operation of the system. Theoretically, unconditional prices of natural resources can be calculated as prices in a closed model of ecological balance as was done with product prices in turnpike models. I realize that, unlike with turnpike models, finding unconditional prices for natural resources is a much more formidable task. Nevertheless, it can be discussed in a theoretical framework.

Taking all this into account, let us examine the empirical aspects of growth measurement when stable valuations are lacking. The cause of many misconceptions that arise in this respect seem to lie in substituting growth measurements for measurement of development.

Note, first of all, that the prevalent approach to growth measurement is based on calculating national income given solely in terms of essential (material) parameters. Furthermore, the working premise (made explicitly or implicitly) is that the prices used to calculate national income exhibit sufficient stability over time. This indicator is quite adequate for the task of measuring economic growth. However, if price gaps resulting from the system's dynamics are taken into account, we see that this indicator fails to incorporate respective shifts in the system (whether in the past or in the future).

Paradoxes arising in this connection stem from trying to construct a growth index between two sufficiently distant time periods based on the same fixed price prevailing at the starting point (especially at a time when significant changes in the structure of both output and prices have taken place). These paradoxes are widely used by political leaders in authoritarian countries to prove the superiority of their system. By varying the base year and its current prices, the leader can achieve a significant improvement in such vital indicators as economic growth.

This situation occurred in the USSR in the 1930s. At that time the country initiated a rapid industrialization program. The so-called Gerschenkron's effect (Gerschenkron, 1951) manifests itself in the fact that the valuation of the economy varies significantly depending on the base point and corresponding weights (constant prices), that is, on whether we choose the initial or the more recent period as the base point of our calculations. Gerschenkron's findings revealed that the growth index for Soviet machine manufacturing for the period between 1928 and 1937 varies significantly depending upon our choice of the base year. This disparity is a result of a drastic change in the structure of output that took place during this period. Restructuring manifested itself in an increasing role of machinery in total output amplified by the fact that the per-unit costs in machine manufacturing between 1928 — when machine manufacturing was in its embryonic stage — and 1937 decreased substantially.

The significance of this paradox in analyzing growth rates of military expenditures was shown in the work of Abraham Becker (1981). Inconsistencies were shown to be most glaring when significant shifts in the structure of military output take place (for instance, a sharp increase in the output of new kinds of weapons).

Taking this into account, the crucial problem faced by economic science in measuring growth rates between two sufficiently distant points of time is measuring economic development. Only in extreme cases when the criterion of development is fixed and all its links with the initial state are known can we attempt to measure rates of growth at a global or even a local level.

Meanwhile, from my reading of U.S. literature and my conversations with a leading U.S. specialist in the field of growth measurement, Irving Kravis from the University of Pennsylvania, I conclude economists do not realize the inadequacy of the tools they are using to evaluate economic performance. They try to measure development by substituting growth for development. This is the source of confusion. These considerations corroborate the statement made in the Introduction that Francis Seton did not succeed, at least for the general case, in finding invariant prices to measure national income outside time and space.

Seton's model for finding the unknown prices is represented in the following way (the model below represents an adapted, but not the most concise, version of the model presented in Chapter 6 of his book [p. 25]).

$$\sum_{j} x_{i}^{j} + \sum_{s} y_{i}^{s} = z_{i}$$

$$\sum_{j} w_{r}^{j} = w_{r}$$

$$\sum_{j} \tau^{j} = \tau$$

where

x_{i}^{j} = sector i's deliveries to sector j,
y_{i}^{s} = sector i's deliveries to end-use s,
w_{r}^{j} = factor r's absorption in the product of sector j, and
τ = residual surpluses and taxes.

This equilibrium representation of economics is conducive to statistical analysis, which Seton undertook on a large scale. Relative to equilibrium, Seton actually seeks to find conditional prices, which, being conjugate parameters, incorporate not only the matrix of technological coefficients (included indirectly via the deliveries to sectors and from the sectors) but also all the constraints on the final products and on the initial factors of production. As I have shown in Chapter 6, this equilibrium model is easily transformed into an optimal one. The latter brings to light certain assumptions hidden in the equilibrium model. An optimality representation reveals Seton's assumption of the existence of a universal model (for all people and at all time), incorporating a unique criterion of optimality in the form of national income under certain constraints on natural resources, capital, and labor. Modulations of the various components of the model can be measured by the change in the value of the criterion or via the valuation of the constraints (resource prices) or the first derivatives of this criterion (prices of final products).

The indeterministic nature of the economy precludes the existence of such a unique model of a system's dynamics. The general case calls for the measurement of development, which, in turn, requires the aforementioned unconditional valuation of material parameters as well as the introduction of relational parameters and their respective valuations. Seton's research is highly relevant provided growth can be measured within the framework of a unique model of optimality.

NOTES

1. I propose a similar definition of the general notion of the potential of an individual as the relationship between his desires and his abilities (one

manifestation of this is the relationship between ambition and ammunition). I have already explained what I mean by desires; the term abilities incorporates such things as personal drive (this is analogous to the resources at one's disposal), various talents (similar to the system's ability to acquire resources, transform them into products, and finally realize the product), intelligence (analogous to the means of achieving an integral balance between desires and abilities), and will power (this corresponds to the ability of implementing decisions into actions).

Thus, productivity of an individual is a function of many variables. Therefore, highly productive individuals are a relatively rare occurrence for they must possess all of these qualities. We know, however, that gifted individuals are extremely valuable to society. At the same time, other traits of their personality may forbid them from realizing their abilities. This fact encouraged various social thinkers to search for and devise all sorts of supports to help these people, including such strong incentives as material rewards, prestige, and patronage from private and government sources.

2. Note in passing that as soon as God nominated Abram to be the patriarch of the Jewish people, He gave him the name Abraham (Genesis, 17:5). At the same time, Bergson who is regarded as the patriarch of Sovietology is called Abram. I believe that in view of the great accomplishments of Abram Bergson, the international conference of Sovietologists must decree that he be called Abraham. Meanwhile, there are some obstacles in the way because the well-known Sovietologist Becker yet has the name Abraham.

3. It was shown in Part III that an optimization approach to growth measurement is just as valid for market economies as it is for planned economies. It was assumed that market equilibrium prices approximate optimal prices. Equilibrium prices from the past are revealed by statistical data (corrections being made in case of disequilibrium states), and future prices can be found in the terms of contracts. In a sense, contracts resemble plans in an attempt to stipulate a coordinated program of future development.

10

Macroeconomy II: Measuring Economic Development in Kuznets's *Modern Economic Growth*

The subject matter discussed in Simon Kuznets's book *Modern Economic Growth* (1966) based on a huge amount of statistical data shows vividly the advances and the obstacles to measuring macroeconomic development. It serves as an excellent illustration of the methodology of measuring economic potential expounded in the previous chapter.

MEASURING MODERN ECONOMIC GROWTH IS MEASURING THE COUNTRY'S POTENTIAL

The book's principal aim is the problem of measuring economic development on the basis of extensive statistical data. The author contends that he is dealing with economic growth. However, within the framework of the scheme outlined above, the problem Kuznets is dealing with is really that of economic development rather than economic growth. The major thrust of Kuznets's philosophy is to elaborate local criteria of measuring the potential for future development. This idea led him to reject the traditional criterion of economic growth as growth of per capita income and instead measure it using a set of independent indicators. In other words his description of modern economic growth incorporates not only material parameters but also relational ones. In addition, Kuznets attempts to link all these parameters and mechanisms together, that is, to analyze their interdependence.

Before we discuss Kuznets's approach let us note that modern economic growth is viewed by the author both in space and in time. In other words his analysis involves both the comparison of different countries with each other as well as the dynamics of the growth of a single country (pp. 23–24). For the sake of simplicity, our discussion shall be limited to only one aspect of the problem — measurement of growth over time.

According to Kuznets, the basic and the most essential aspect of the problem is the prevailing system of values governing a given economic epoch. Kuznets associates modern economic growth with an epoch that is characterized by the emphasis on secularism, egalitarianism, and nationalism. Indeed, for a society with a different value orientation, for example, for a religious society, the problem of modern economic growth with its national income accounting methods is irrelevant (p. 21).

The author also points out that different economic epochs traversed by a given country are compared historically in terms of the existing system of values. The past is evaluated in terms of the values currently prevailing. This does not exclude the possibility of "using different systems of notions for different economic epochs" or using a system of values contemporary with the system under consideration (p. 23).

As I mentioned, Kuznets suggests using a set of indicators to measure modern economic growth rather than a single criterion traditionally represented by the national income or per-capita income. The point is that "no single criterion of distinction can fully exhaust the determinants of a complex process" (p. 19).

In constructing this set of indicators, the author warns us that this set is not exhaustive, for our knowledge of the structure of the system is incomplete (p. 25). In mathematical terms, this means that the set of singular points characterizing the current state of the economy is not sufficient to determine the structure of the entire set of which they are a part. In principal, such a set of singular points could be constructed, but even then "they would remain nonadditive unless we could assign proper weights to them, which we cannot do now" (p. 26).

So, how does Kuznets propose to integrate the set of indicators characterizing modern economic growth or, to use my terminology, the potential of the system?

Individual material parameters, assuming they are homogeneous and can be represented in real terms, pose no problem. To put it more simply, these are cases where material parameters, such as population, can be quantified in physical units. At the same time, there exist material parameters that do not lend themselves to such a representation because they are composed of heterogeneous elements. Such cases (for instance, national product) involve commeasurement of various goods and services among themselves. Kuznets realizes that the existing method of measuring national product in current prices is imperfect, but he does not provide us with any alternative procedure of deriving the values of these parameters.

Kuznets does not even attempt to elaborate procedures of evaluating relational parameters, as well as the economy's mechanism of operation, as independent variables. However, he does stress one very important point when he writes that the evaluation of an overall economic growth should not be based on a number of individual relational parameters, no matter how significant they are, for example, the percentage of technologically advanced sectors in the whole output (pp. 24–25).

In view of all the difficulties in measuring individual components of modern economic growth, aggravated by the impossibility of avoiding this task by concentrating on the "leading" sectors of the economy, Kuznets makes an attempt to measure an overall level of modern economic growth. By introducing a number of conditions in measuring modern economic growth — first and foremost, the minimal duration of the period under consideration — the author manages to avoid many pitfalls associated with general representations of the economic system. His "minimum of 30–40 years" is required to reveal short-term fluctuations in the magnitudes of parameters used in measuring modern economic growth (p. 27).

This choice for the minimum length of the period allows Kuznets to analyze the problem in a qualitative way or, to be more precise, to establish the epoch (using Kuznets's terminology) to which a given economic system belongs. Parameters determining modern economic growth are interdependent, and, when viewed from a long-term perspective, they are found to be strongly correlated. Therefore, we should expect all of them to move in the same direction: growth for some and decline for others.

As a first approximation, we can establish the type of a system by the direction of change of the relevant parameters without going into a more detailed analysis of the relative rates of change of each individual parameter. Kuznets gives statistical data to support the ideas advanced in the book. For instance, in most of the developed countries undergoing modern economic growth, growth of per-capita product fluctuated between 15 percent and 30 percent per decade while population grew at about 10 percent per decade. This means an "increase in aggregate product ranging from 20 to close to 50 percent per decade" (p. 490).

COMPONENTS OF MODERN ECONOMIC GROWTH

The set of components used by Kuznets to measure modern economic growth includes primarily the material parameters. The author is interested in their magnitude per se (for instance, size of the population or absolute level of per capita income), but, as I mentioned above, the problem of measuring the values of these parameters as parameters contributing to the potential of the system is not discussed.

The main emphasis of the book is on the structural parameters affecting modern economic growth. The methodological significance of this approach lies in its utilization of the principle by which structural parameters belonging to different steps of the sequence are formed, in following this chain to some final state represented by a material parameter, and in treating these structural parameters as multidimensional entities.

A structural parameter such as population growth — step number one — should serve as a good example of the use of this methodological principle. Excluding such things as migrations, population growth is determined by such structural parameters as birth and death rates — step number two: a country experiences population growth if its birth rate exceeds its death rate. Growth can be achieved via higher birth rates, lower death rates, or both. For modern economic growth lower death rate was the major factor responsible for growth of the population. The death rate by age, in turn, forms a peculiar U-shaped curve — this is step number three. Lower death rates are due for the most part to the change in the curve's shape: infant mortality has decreased more than mortality in the more advanced age groups. Finally, we arrive at the material factors responsible for lower infant mortality. They are directly connected with advances in the medical field: preventive measures and treatment of infectious, respiratory, and digestive diseases most prevalent among the young (medical science has been less successful in treating such ailments as cancer and cardiovascular disease, more common among adults) (p. 45).

Structural parameters associated with population were analyzed from different points of view: population dynamics over time for the country as a whole as well as for rural and urban areas separately. This evaluation produced a better picture of the reasons for population growth. Thus, as late as the beginning of the twentieth century, death rates among children and old people living in cities were much higher than death rates among these same age groups living in rural areas (p. 46).

Another structural parameter subjected to a thorough analysis was product per capita — step number one. Growth in per capita product was due to the rate of growth of production exceeding that of the population — step number two. Growth of production was, in turn, achieved by increasing productivity of labor and capital (including restoring capital) — step number three.

Kuznets then proceeds to analyze a number of structural parameters characterizing changes in the structure of the economy and capable of revealing information about the changes in the structural parameters appearing at step three. This is accomplished primarily by examining the interrelations among the leading sectors of the economy: agriculture, industry (here understood to include construction and transportation), and service — this represents step number four. In revealing these structural changes characteristic of modern economic growth, the author goes on to analyze individual sectors of the economy, especially its leading sector — industry. He uncovers connections among the leading branches of industry that include mining, output of gas and energy, water supply, manufacturing industries, construction, transportation, and communications — this represents step number five. In the end, this sequence of steps leads to such material parameters as technological innovations.

One manifestation of Kuznets's multidimensional approach is his analysis of the structure of the economy from two different perspectives: relative volumes of output produced by each of the sectors and relative size of the labor force attached to each sector. When viewed in conjunction with each other, a deeper understanding of the dynamics of economic development was achieved.

Using a broad statistical base, Kuznets examines in detail structural parameters related to output and income distribution, output allocation (for personal consumption, government consumption, or capital investment), and international economic issues. Nevertheless, the author notes that a number of structural parameters significant for modern economic growth could not be verified because "they do not lend themselves to quantitative measurement." One such structural parameter is "accumulation of useful knowledge and the rate of technological change" (p. 490).

Characteristic parameters are less developed in Kuznets's work. As I have already mentioned, the main thrust of the book is on the structural parameters that are subjected to a systematic analysis with an attempt to assign weights to them (in other words, to discover the impact of a given parameter on the leading indicators of modern economic growth). Occasionally characteristic parameters do come into play, but they are not analyzed systematically (often as a by-product of the main discussion), and no attempt to assign weights to them is made. This does not mean that the characteristic parameters mentioned in the book are treated as insignificant. On the contrary, although they are assumed to be important, the problem of assigning weights to them, as well as the problem of finding an economic index common to a given epoch, is avoided.

One vital characteristic parameter regarded as having an explicit impact on modern economic growth is mobility. Increasing mobility is characteristic of all countries undergoing modern economic growth. Mobility permeates every aspect of the economy. It can be seen in a rapidly changing structure of demand for the final goods, in the structure of the economic sectors and their branches, in shifts in the structure of the labor force (not only in terms of its allocation by sectors, but also by status and occupation), in the type of business organization prevailing in society (a shift from small firms, especially common in agriculture, to intermediate and large-scale firms), and in the relative share of foreign trade to domestic output.

Another characteristic parameter examined by Kuznets is parallelism. For instance, one aspect of modern economic growth that seems important to the author is the parallel growth of both production and population. This example was discussed previously.

One more characteristic parameter, also touched upon in my previous example, is the stability (stationary nature) of some structural parameters. This characteristic parameter is discussed in connection with another major indicator — the growth of national product (here meant in the broad sense of the word):

> The second feature of modern aggregative growth . . . is the relative steadiness of its rate over time — compared with the variability of such rates in premodern times. This conclusion is highly probable, particularly for rates of growth of total product: viewed in, say, decadal units, the very steadiness of modern population growth, freed from the catastrophic impacts of epidemics and the fluctuations in birth rates associated with them, would make for a greater stability in the rates of growth of total product. Other contributing factors are: technological progress in agriculture, which reduces the chance of crop failures; improvements in transportation, which permit the widening of sources of supply; the decline in the share in total product of agriculture, and the increase in the shares of other industries, technologically more controllable — all basic features of modern growth (p. 70).

Besides characteristic parameters having a positive impact on development in the modern epoch, Kuznets examines characteristic parameters that interfere with development. The purpose of introducing these "associated" characteristics, discussed rather briefly, is to isolate countries that, in spite of having a high per capita income, do not belong to the family of countries experiencing modern economic growth. The author distinguishes two such characteristic parameters: political sovereignty and monoproducts (pp. 488–89).

Finally, the author mentions another very important component of modern economic growth — the political mechanism of the country. This characteristic is not applied to developed Western countries but is used primarily to eliminate communist countries that have many features, such as utilization of "trans-national stock of useful knowledge," which would otherwise place them among countries with modern economic growth (p. 358).

I have to agree with Kuznets's far-sighted conclusions concerning the placement of communist countries in a separate category. Indeed, developments in the last 20 years (remember that the book came out in 1965) have shown communist countries, and the USSR especially, to exhibit tendencies opposite to those characteristic of modern economic growth. This manifests itself in a rise in death rates among children and middle-aged men. On top of these calamities, starting at the beginning of the 1980s the USSR has exhibited a marked decline not only in the rates of growth of output but also in output in absolute terms. In my opinion, this deterioration is not a result of transitory short-term fluctuations or economic cycles; it is a result of a long and steady process of exhausting the system. It remains unclear whether the former Soviet republics will be able to stop this process — to promote real (rather than verbal) long-term development on the level of modern economic growth without making major changes in the political structure of the system (Katsenelinboigen, 1990).

Other considerations besides the ones advanced by Kuznets enter into play in excluding communist countries from the family of countries enjoying modern economic growth. Broadening the scope of our analysis of communist countries to include structural parameters characterizing their economies would also lead us to conclude that these countries face the danger of stagnation. The said parameters have to do with the structure of the output and with other factors responsible for economic growth. The output of communist countries was structured to accommodate military needs. Peacetime goods and services were no more than a by-product of a militarized economy. This set-up implies not only that the greater share of output is channeled directly into the military (production of weapons and maintaining the army) but also, and this is really crucial, that output policy was geared toward quick convertibility to wartime needs (for instance, instead of producing light-weight wheel tractors, heavy-duty caterpillar tractors, which are easily converted into tanks, were produced in large quantities) and the priority of increased output of versatile basic commodities such as steel, electric energy, coal, and oil. This policy, of course, exerted a strong impact on the structure of the economy. Until very recently, principal growth occurred in the industrial sector while agricultural and service sectors, if they grew at all, grew at a very irregular pace definitely lagging behind the demands of the population.

The factors affecting economic development of the USSR was changed primarily through an extensive way of growth, that is, by employing a greater labor force and more capital, which sooner or later had to exhaust the country's resources. Mod-ern economic growth is characterized by intensive growth, that is, growth achieved via such factors as technological and organizational progress.

MODERN ECONOMIC GROWTH AND THE CONCEPTION OF A POTENTIAL OF A COUNTRY

I conclude our discussion with the following question, "Does modern economic growth promote the growth or the decline of the country's potential?" The answer to this seemingly simple question turns out to be not so simple if we view modern economic growth in a more general context of human development. Other considerations create serious doubts whether modern economic growth really increases the potential of the nation and of the human race in general. There are good reasons to believe that modern economic growth creates greater tension in the potential of different nations and may even lead to an all-destructive conflict. Kuznets gives an interesting analysis of many such reasons, but I discuss just one of them (p. 500).

It seems that in the past the aggressive side of human nature found other ways to express itself. Strong religious beliefs used to bind people together in initiating wars. Even today this phenomenon is on the rise in the Middle East. In the age of strong family ties, vendetta was common, and strong local ties led to one neighborhood fighting another.

Hitler and Stalin have demonstrated that nationalism (however justified), an attempt to declare one nation superior to other nations thereby granting it the right to rule the world, became one of the major causes of international conflicts in the twentieth century.

The problem of increasing economic potential by means of technological progress has been a subject of debate for many centuries. Reducing its danger to the growth of nationalism would be an undue oversimplification. The dangers of technological progress as the basis of economic well-being seem to be much more profound. The phenomenon of technological progress reflects, in Kuznets's opinion, the eternal cravings of mankind (p. 501).

I agree that many people do dream of these things, but let us not forget that some sophisticated civilizations understood the dangers of technological progress. Oversimplification of this old problem and an inability to understand that it involves an act of belief can only be attributed to the arrogance of Western thought, which considers it absolutely plain and natural, or very human, to aspire toward greater happiness primarily by means of technological progress. We have no criterion to judge the validity of these claims because our information is incomplete. Unfortunately, Kuznets pays his dues to this Western arrogance and considers the "bare minimum needed to hold body and soul together" a view limited to religious asceticism (p. 103).

We should not be so hasty to pronounce judgment of culture — whether it be introvert or extrovert culture having technological

progress.[1] Elements common to the aspirations of all should not be reduced to the unification of the means (which may themselves pose a considerable threat) through which human desires can be fulfilled.

In summary, we could state that Simon Kuznets's book represents a nontrivial attempt to flee the bounds of the classical approach to measuring economic growth based on national income. The author deserves much credit for introducing various relational parameters that, in conjunction with material parameters, determine the economic potential of the country. Moreover, the entire analysis is based on a vast amount of data.

At the same time, Kuznets's rather narrow conceptual framework kept him from recognizing more clearly the difference between economic growth and economic development. As a result, he never embarked upon the new avenues of dealing with the problems arising in trying to measure economic development. It seems to me these problems really entail the creation of a new system of valuations of parameters incorporated into economic development.

NOTE

1. In my work on the interpretation of the Torah (1989), I distinguished between extrovert cultures with and without technological progress. In the latter case, extroversion was achieved by means of greater intensity of work and, more important, through better organization.

11

Macroeconomy III: Convertibility of an Economic System

The problem of the convertibility of an economy from a peace to a war mode, and vice versa, stands by itself as a distinct problem. Works that I am familiar with fail to treat the problem of economic convertibility as part of the general problem of creating economic potential although the two problems share much in common. But before discussing the essence of this particular problem, it is necessary to define the notions of peace and war economies within the scheme of a multifaceted methodology of the systems approach, that is, to consider functional, structural, process, and genetic aspects of the problem.

DEFINITIONS OF WAR AND PEACE ECONOMIES

Let us first imagine two ideal extreme situations. The first one is a peaceful tribe of herders, isolated in remote mountains. The notion of a peace economy fully describes this tribe, with its war expenditures being equal to zero. Now let us consider another extreme, a country in a state of deadly war with its enemy, for example, the USSR in the period of the war with Nazi Germany. In order to describe this country, we can fully use the notion of a war economy. Practically all expenditures made at the time, from food for the workers to the production of tanks, were war related. Between these two extreme states lies a full spectrum reflecting the degree of militarization of the economy, defined in the following way.

Let us draft two plans of development for a country, taking into account the prevailing constraints upon human and natural resources and scientific and technological know-how (at the international scale). One plan will be geared toward current needs (in an extreme case, toward the ideal of peaceful development); and the other one, toward another ultimate goal of winning the war. Let us now take the magnitudes of output parameters derived from plan 2 and impose them as constraints (with a less than or equal sign) on

plan 1. Now we shall devise a new version of the plan geared toward current needs but subject to the new constraints. The difference between the value of the objective function of the two drafts represents military expenditures, that is, the greater the difference, the lower the military expenditures. Roughly speaking, military expenditures in a military economy equal all the country's expenditures minus the production cost of consumer goods and services, which would be cut back in the event of a war.

The expounded scheme represents a functional approach to measuring the degree of militarization of the economy. To develop this point further, let us look structurally at military expenditures and divide them into three parts: direct, conjugated, and indirect expenditures. The category of direct expenditures includes direct outlays for arms production, construction of military installations, and payments to military personnel.

Conjugated military expenditures are the additional costs of peacetime production called for by the need to sustain a militarized economy. For example, if a country were at peace, the development of its coal industry would have been primarily based on the activation of mines with deposits of high quality coal and good mining conditions, as well as adequate means of transporting the coal to consumers. But suppose that a certain portion of the country is militarily vulnerable. This vulnerability has to do with the lack of segments of the population loyal to the government and able to participate in a resistance movement in case of foreign occupation in the region. In case of war the enemy can easily march through this region and occupy it, and thus the problem of creating these population segments appears. The mine workers, being physically strong and enduring and whose high salary makes them relatively more loyal to the political regime, are one such population segment. If some coal deposits are found in a vulnerable region, then, in spite of the low quality of the coal, the susceptibility of the area to floods, and the resulting difficulties in mining, the mining of coal would be organized in the region in time of peace because of military considerations. The extra expenses the country carries because of mining this kind of coal would be conjugated military expenditures. They would not be accounted for by any military board, and the cost would flow into the regular financial streams among civilian businesses.

By indirect expenditures, I understand the costs associated with the nation's general militaristic orientation, that is, with the policy of maintaining an economic potential for the satisfaction of defense needs. The military-economic potential includes such items as steel, oil, coal, steel cutting equipment, and all kinds of transportation equipment and facilities. In this case, the objective of steel production, for instance, is a greater output of the same. These expenditures are extremely difficult to calculate, but they form a huge part of the defense burden.

From the standpoint of the process, what defines a military economy is that the government places orders not only on the final goods required to cover direct military expenditures but also on the intermediate goods that are requisite for the potential growth of the military might. The latter entails primarily conjugate and indirect military expenditures.

From the standpoint of genesis, a military economy exhibits orthodoxy in the realm of foreign policy and a long history of expansionistic tendencies. This aspect is helpful when establishing the duration of the time period to examine the status of the country's economy, for instance, in distinguishing between a short-range burst of military output and a long-term orientation of the country toward military ends.

Everything stated above is nothing more than logical exercises. In order to show how the suggested measures can be verified, it is necessary to conduct a reasonably serious empirical analysis of the different types of economic systems. In case of a favorable result, the tests will show how we can really use these measures to make judgments about different types of economies.

Intuitively, it appears to me that, within the elaborated scheme, the economy of the former Soviet Union was a militarized one and the U.S. economy is probably peaceful, but this requires further statistical tests.[1]

As far as further discussion is concerned, I will talk about the preparation of the economy to convert itself from peace to war and, conversely, from war to peace. I will be considering these problems largely by using the materials published in the West because in the former Soviet Union, China, and other socialist nations, the given problem is rarely discussed in the open press, even in relation to the countries themselves.

From the point of view of the role of the indeterministic methods in solving economic problems, which constitutes the subject of this work, the more interesting discussions are done in relation to the transformation of a peace economy to a war economy. The problem of the transformation of a war economy to a peace economy demands further discussion.

When I talk about the transformation of one type of economy to another, I mean, primarily, the preparation for this kind of transformation. Although the analysis of the actual period of the actual transformation is also important, I do not deal with it here. I can only mention that there are interesting problems in the actual process of transformation because it is important to conduct this transformation in such a way as to maximize the efficiency of future development. For example, precisely in the period of such a transformation it is easier to renew the capital stock in the economy. The point is that such a renewal is connected with huge expenditures, which are difficult to make even in peacetime because the populace may not be ready to accept the decrease in the standard of living that may be required. In times of war, however, people are used to making sacrifices. In view of the above, right after a war, when war expenses decline sharply, it may be possible to ration the saved resources in order to combine the growth of the standard of living of the populace with the renewal of major funds.

From the point of view of the ability of a system to convert to a peacetime footing, the problem of economic indeterminism may be studied in relation to various levels of development of war production. It may be studied in application to the complete war economy of a country involved in a deadly confrontation with the enemy, to the war or militarized economy of a country (which also conducts local wars) in time of peace, and to a peaceful country that has a significant military sector.

As far as I know, questions dealing with the convertibility of a war economy to a peace economy have not been methodologically considered as acutely as the convertibility of a peace economy into a war economy. Probably the process of preparation to transform a peace economy into a war economy is not symmetric to the converse process, the preparation to transform a system from a war economy to a peace economy.

This asymmetry in dealing with the problem of convertibility is not unfounded. It is determined by the difference in importance of reaching a given goal: military victory or improved standard of living (these two goals adequately reflect the two types of economies). In times of war, when the war economy develops ideally, all resources are concentrated on achieving victory, not on what the country is going to do once victory has been achieved. The efforts that are required for creating an effective war economy are so large that to think about how to provide at the same time for transition to a peace economy, how to combine in the best way the solutions to the problems of economic development in war and peacetime, especially if it is somewhat necessary to sacrifice military strength for this, appears to be quite difficult.

A similar situation may be observed in a country in which a war economy was created in time of peace, especially because of an expansionistic foreign policy. The difficulties of transforming this type of country to a peace economy have primarily political rather than economic constraints. I will discuss this question in a separate section, using the example of the former USSR.

The situation is different for a country that is not at war, and not under any immediate military threat, at least according to its leaders. Here war production in peacetime may also occur, and the problems of convertibility to peaceful production reduce to significant dilemmas in the development of the country because there are various possibilities of political and economic development of the country and various military doctrines (for example, defensive and offensive) that may affect one another. In such conditions, the problem of convertibility of war production to peace production is solved somewhat differently. I consider these problems in a separate section, using the example of the United States.

Our discussion of the problem of convertibility will be structured in the following way: first, I will examine the convertibility of a peace economy to a war economy, and then I will discuss different cases of convertibility of war production to peace production, in war and peace economies.

CONVERSION OF A PEACE ECONOMY TO A MILITARY ECONOMY

In the present context, the deterministic approach represents a program of transition from a peace economy to a war economy. The design of such a program may, in turn, proceed along two lines. The first has to do with a holistic examination of the peace economy and its transformation into an entirely military one. This kind of analysis can be performed at the level of individual industries using the methodology of input-output tables.

The second approach examines the potential to mobilize resources within a few especially critical sectors of the economy in terms of expanding these sectors' production and providing reserves of goods they produce. A number of works dealing with the transformation from a peace to a war economy within one single area were published in the open press. For example, there are significant publications on the United States' mobilizing ability in the field of mineral resources (King, 1981). There are references to the research conducted in 1977 by the U.S. Department of Defense and the Office of Management and Budget the research by Arthur D. Little (1974). The first project aimed to determine the strength of the airline industry.

I will discuss more thoroughly the indeterministic approach toward uncovering the potential of converting a peace economy to wartime. It involves finding the economy's predisposition to perform in the event of a war. I will limit myself to examining this problem without explicitly accounting for direct military spending at a time of peace. This factor will be discussed in the next two sections, where the problem of converting a war economy to a peace economy is examined.

Taking these comments into account, the work of Irving Leveson and James Wheeler (1978) deserves special attention. Using a wide range of statistical material, this work looks at the economic and social problems of transforming the United States from a peace economy to a war economy. The authors call these problems "mobilizing potentials of USA." They also picked an interesting angle for the analysis of this issue, namely the influence of the current and future social and economic changes in the United States upon the mobilizing potential of the country. This work extensively discusses many parameters that define the predisposition of the peace economy of the United States to switch to a wartime mode. The change in the magnitudes of these parameters may affect the mobilizing potential in a negative as well as in a positive way.

Let us look at what the authors have done in developing the concept of potential, given the limitations mentioned above. They have suggested the following classification of the various parameters that define the mobilizing possibilities. First, the parameters are divided into two groups — acting inside and outside the state.

In analyzing the inside position, the authors specified human resources, macroeconomic factors, and social factors. In discussing the outside situation, they analyzed such factors as transportation, export, and import. The discussion of the above factors was accompanied by the analysis of their various characteristics.

I will discuss this work and classify the issues it raises from the point of view of my concept of potential, that is, the notion of material and relational parameters (the latter are split into structural and characteristic ones), as well as the mechanisms of performance. Of course, all these elements of the potential are interconnected, but I shall group them depending on which elements the authors emphasize.

First, the authors consider the role of material parameters dealing with resources and people. The authors discuss such material parameters as the qualification of workers in the service sector and they show that the current level of development of the service sector, which is adopting electronics as well as methods of mass production on a wide scale, creates a significant number of people possessing managerial experience as well as a skilled profession.

The authors analyze in detail the multitude of structural parameters, which reflect the country's readiness to mobilize. Most prominent among the structural parameters are the structure of labor, such as the sex ratio among workers in the various branches of the economy. The relative number of the workers in different sectors represents another structural parameter. The greater is the share of workers in the production sector as opposed to the service sector, the greater is the mobilizing ability of the country. Meanwhile, the workers' skill level in the service sector is growing and, therefore, the growth of the service sector does not have such a negative effect on the country's mobilizing abilities.

The way the authors analyze the structural parameters that deal with the relations between the United States and other countries is quite interesting. They look at the structure of the imported goods in terms of agricultural/manufactured classification: the less is the country's import of agricultural products, the greater, in principle, are its mobilizing abilities. If a country does not produce certain agricultural goods, then reserves are required. In the United States this relates to a very small number of products, for example, natural rubber, some herbs.

The authors also pay a great deal of attention to the characteristic parameters, which must be considered when calculating the country's mobilization readiness. For instance, they assign a significant value to such parameter as cyclicity. The value of this parameter is determined by the mobilizing ability, depending on whether the mobilizing period occurs during the time when the production capacities are free or strained. It appears that the authors, in principle, evaluate this parameter positively because cyclicity increases the ability of one country to compensate for its difficulties by using the cyclical situation in another country.

A large number of characteristic parameters affecting mobilizing ability are analyzed by the authors in some detail when they consider foreign relations between the United States and other nations. Here, such parameters as openness, vulnerability, versatility, and maturity are examined. For example, openness comes into play primarily for understanding the nation's dependence on its export to other nations (for example, acquiring much revenue from exported goods, and especially capital) as well as imports from these nations.

It should be noted that the authors pay attention to the functional connection between structural and characteristic parameters, which ultimately allows them to make the connection with the material parameter representing the ultimate objective, namely the mobilizing ability. For example, the structural parameter that reflects such material parameters as knowledge and income may be linked with such characteristic parameter as mobility. Here is what the authors write about this: "There is a strong tendency for geographic mobility to be greater among persons with greater education, presumably because they tend to acquire more information and because the costs of moving tend to be low relative to income" (p. 16).

A number of other components of the system's potential such as the existence of rules (laws) within the economic system, and the ability to choose among different courses of actions depending on a situation, are analyzed in the work in question. This also applies to the methods of system's operation, say centralized versus market. In addition, these components of the potential are often linked with the structural and characteristic parameters because their importance, by and large, defines the effectiveness of the respective rules.

The authors note, for instance, that the changes in the attitude of the masses in the United States, which decreased the value of such a parameter as trust in the government and the leaders, may have a strongly negative effect on the mobilizing abilities of the country. Mobilization may call for an increase in the central power (extreme cases may be exceptions) (pp. 28–30). More specifically, the danger that it may be quite difficult for the government, given the aforementioned environment, to provide on short notice the transformation of the economy to military needs affects the structural parameters because it shifts the structure of the reserves: instead of reserves of raw materials, it needs to develop reserves of finished goods (p. 30). This brings the authors to consider the general aspects of the interplay between centralized and decentralized methods of operation of economic systems.

While the above research is highly valuable, primarily because it examines a variety of parameters comprising the potential of the system, and quantitatively verifies each of them statistically, I would like to note its limitations. First, the work does not give quantitative values of the various parameters. Second, the discussion of the problem is presented outside the general concept of economic development — more as a specific problem of convertibility of an economy, rather than the holistic concept of creation of economic potential.

ADJUSTMENT OF A WAR ECONOMY TO A PEACE ECONOMY

If the country's idea of development is based on expansionism, then the strengthening of military power becomes top priority. Such an economy is geared toward military goals, and the problem of convertibility of military production is reduced to introducing certain adjustments having to do with peace-time output.

These adjustments are stipulated by the fact that in time of peace quite a serious problem arises: the use of capacities primarily intended for military needs arises because the need for military goods that can be produced by these capacities is less than at a time of war, and it is impossible to use these capacities for expanding and renewing war-oriented production. The need to adjust existing capacities to peace production, in such a situation, is dictated by at least two reasons. First, it is important that these capacities are not idle, if only because the workers would lose the experience. Second, it is necessary to provide the populace with consumer goods in order to create a material incentive to work. But, I want to reiterate, that the solution to these problems is nothing more than the adjustment of military production for peacetime needs.

Let us look at the above case using the example of the former USSR (Katsenelinboigen, 1990). The Soviet system, continuing by and large the traditions of Russia, from the very beginning of industrialization built its economy to maximize its readiness for war — in the end, an offensive war.[2] The goal was sufficiently clear and it claimed all the resources. The structure of the government having a double hierarchy — economic and party — was created, which planned the country's development and carried its plans out in administrative ways (by watchdogging the people to fulfill the plans).

I believe that the thinking of the Soviet leaders in this sense was clearly combinational and, from their point of view, quite successful. Russia turned into a superpower with a modern army and state-of-the-art weapons. The Soviet empire grew significantly and expanded its sphere of influence in many countries.

However, such a successful combinational game was fraught with the danger of a Pyrrhic victory — the inevitable companion of a combinational style when the system's potential is forgotten. It appears that this is exactly what took place. I believe that the slowing down of heavy industry, which started at the end of the 1970s, and the deterioration of agriculture and light industry are the results of exhausting the potential of the country. In other words, the current decline of economic activity and military might is not the consequence of some normal fluctuations but is the result of mutilation of the system by the militarized economy and the accompanying structure of the government.

The profound changes in the potential of the country over the last 70 years inevitably led to stagnation. These changes concern the material parameters, as well as the parameters of relations (structural and characteristic), and the structure of government.

The negative changes in the values of material parameters are widely known. They are manifest in the following. There appeared a huge gap between the needs of the country and the available natural and human resources, as well as capital. The material parameters of the country proved to be inadequate for preserving its status as a world power while meeting current military demands and the people's standard of living.

Furthermore, the quality of products produced in the USSR was low and very often did not meet world standards. The low quality of the products was the consequence of either the desire to increase the quantity of production regardless of cost, frequently disrupting the technology of production in the process, or cost inefficiency of additional spending on improving the product quality because certain kind of quality improvements important for peacetime are not vital for military production.

Let us consider in more detail which structural parameters were included in the potential of the former Soviet Union and why they eventually led to stagnation.

In general, the structure of output for military purposes and for the people was set up in the USSR in such a way as to increase the output of the products needed for military purposes; the production of consumer goods was looked at as an inevitable evil to be minimized. The structure of the economy was built to correspond to the prescribed structure of output. Thus, heavy industry was emphasized, as opposed to light industry, agriculture, and the service sector.

In the meantime, the structure of consumer goods produced was designed in such a way as to be compatible with the development of heavy industry, that is, the production of those goods that could be produced using the capacities employed directly for military purposes. Hence, there was a significant growth in the production of watches, radios, television sets, refrigerators, and other consumer goods and slow development of production of food products, clothes, and housing and such.

The structure of the production of capital goods also reflects the above priorities. The mass production of tractors as opposed to agricultural chemicals (fertilizers, pesticides, etc.) was set up: the capacities used to produce these chemicals are not as necessary in time of war whereas the tractor capacities provide for substantial production of tanks and trailers for missiles. However, in order to raise the productivity of agriculture, the production of different types of chemicals should have been stressed as well.

As far as tractors, most were heavy machines. Although for agricultural needs the production of light or medium machines (with attachable equipment) would have been preferable because they are cheaper to produce and operate and allow for the conservation of soil, from the point of view of military needs, tanks and heavy trailers were required.

Moreover, the full subordination of the economy to military needs in peacetime had lead to certain imbalances in the economy. For example, difficulties in providing agriculture with spare parts for delivered machines arose. Indeed, from the military point of view, the capacities for the production of tractors should not be oriented toward the production of parts, which last for a long time, or toward the production of significant amounts of spare parts. This is caused by the obvious fact that in time of war the life of a tank is rather short, often being measured in months, whereas a peaceful economy needs tractors with a long life span and a good system for distributing spare parts. In what was called the Soviet Union, the capacities of tractor plants are geared, primarily, toward the production of tanks. This explains the corresponding proportions between the mechanical and the assembly shops at a factory (spare parts are made primarily in mechanical shops) and technological processes oriented toward the production of parts with a relatively short life span. Certainly, in peacetime the economy pays a huge price for this. It is well known that almost half of the tractors in Soviet agriculture were cannibalized because spare parts have been stripped as parts for working tractors. In spite of this, there is still a lack of these parts.

Therefore, the structural parameters of the Soviet economy discussed above bear witness to its complete military orientation. These structural parameters, having at best a positive value from the point of view of attaining military objectives in a limited time, have a negative value from the point of view of the development of a country's potential.[3] Because in the USSR peace production was subordinate to military production, the structure of peace-time output and the structure of the production capacities as a whole were essentially deformed.

As far as the characteristic parameters are concerned, they also have a primarily negative value. Some of these parameters like openness and exhaustion have been analyzed in my book (1990).

Thus, instead of primarily orienting the Soviet economy to effective peacetime production, it was oriented to military production, and the adjustment to peace time conditions demanded minimal resources. All Soviet factories were built so as to ensure the production of military goods and each one had a mobilizing plan for this purpose. Usually factories produced goods that can be directly used for peace or war ends. Here, adjustment of the production during peacetime concerns only the degree of utilization of military capacities for peacetime ends and depends on the international situation and the desire of the leaders to be good to the masses.

Therefore, in the former USSR we recognize a sufficiently sophisticated system of a military economy in peacetime and this system had its own logic of development, although this logic is not always impeccable (Katsenelinboigen, 1978).

As the above noted factors which shaped the Soviet militaristic economy became more pronounced the potential of the country declined abruptly and began to pose a threat to the further growth of the country's military might.

In order to halt the stagnation of the Soviet economy and turn the economic process toward stable development, it was first necessary to challenge the main objective — the growth of military power. Taking extreme cases, there appear the following two alternatives. The first one is to preserve the strength of military power as the leading direction of the country's development while safeguarding the authoritarian regime, that is, without real limitations upon the leader's power. Meanwhile the allowed methods of action support a respite to decrease military spending for a few years in order to use the freed resources primarily for renewing capital stock, improving the quality of goods produced, and ensuring a certain growth in the people's standard of living. Undertaking such a respite could help change the methods of economic performance and be conducive to the country's gradual transformation to a market economy which is otherwise incompatible with a completely militarized economy.

It is feasible to assume that Gorbachev wanted to implement the above mentioned policy of respite. In the frame of Russian history it looks like an attempt to preserve such invariants of the seven-century-long Russian tradition as autocracy and empire. However, in doing so he wanted to follow not the harsh methods employed by Ivan the Terrible, Peter the Great, or Joseph Stalin, but the more flexible strategy practiced by Alexander II and Nicholas II. However for Gorbachev such a policy had a tragic ending.

It is not clear at this point what course of development the new leaders of Russia and other industrial republics (Ukraine, Byelorussia, Kazakhstan) would choose to follow. They will either try to preserve the authoritarian regimes and continue to develop military might but this time, taking Gorbachev's mistakes into account, they will undertake a deep respite accompanied by a forceful restructuring of the militarized economy and new methods of management. The second alternative which they have at their disposal is to create some potential for peaceful development. The allowed methods of action are to orient the structure of goods produced and the existing capacities toward peaceful production, to gradually create a pluralistic democratic (not to be confused with ochlocratic, that is, mobocratic) open system with the separation of power, and to preserve the level of defense sufficient to meet a real threat of attack.

The tradition of Russian and Soviet empires makes me rather pessimistic. Only time will tell which course the leaders of Russia and other former Soviet republics will choose to follow!

CONVERTING MILITARY PRODUCTION TO PEACE PRODUCTION IN A PEACE-ORIENTED ECONOMY

There is a considerable number of works in Western literature dealing with the problem of convertibility of military production into peace production in the nations of the free world, which are assumed to be countries with peaceful economies. I consider some aspects of this problem using primarily the book by Seymour Melman (1974). It gives quite a wide, although somewhat radical, picture of the problems of U.S. convertibility of military production to peacetime needs.

Even with respect to material parameters, however, everything is not this simple in countries with market economies because the expansion of military production is sometimes viewed positively. The paradox of the problem is that some professional views purport that the development of military production in a free society, especially during years of overproduction when it is important to find new sources of demand, may turn out to be quite a strong boost for the economy and, in general, for the growth of the country's potential. Overcoming the deepest economic crisis of the prewar years in the United States by increasing military production, expanding the industry, and raising people's incomes during World War II led to the popular view that the increase in the military budget was good for the country.

The views of Western governments on cutting military production cause much discussion (Udis, 1987). There are radical views that certain circles in the United States are not interested in cutting military spending and transferring funds to peaceful production. There are also opposite views. Congress is interested in this problem, and the government provides financing for appropriate studies by top scientists in the country.

A permanently high level of military production, however, may significantly decrease the economic potential of a country. Military production may, under such conditions, become a parasite, sucking out the juices of a living organism. Moreover, growing military production may increasingly encompass the economy, change its very nature, and eventually even suffocate it. In other words, the expansion of military production may negatively affect the economic potential of a country mainly by decreasing its material parameters.

Military production may also have a highly negative affect upon the structural parameters of the potential, primarily because it changes the structure of the produced goods by channeling valuable resources to the production of goods that will not be used for further production or consumed by the public. Significant changes then occur in the structure of the work force: military production attracts the best workers.

The characteristic parameters of economic potential during a rapid development of military production may also change, and the ones with a negative effect on the potential will dominate. It is sufficient to give an example with such characteristic parameter as openness-closeness. The development of military production may bring closeness to a point at which it begins to play a highly negative role. The secrecy of scientific and technical developments increases, and, therefore, access by scientists and engineers to the newest results is cut abruptly. Military developments and military equipment must be prevented from getting into the hands of undesirable nations.

Additionally, every limitation, if it conflicts with the common interests of the people, inevitably leads to ways of overcoming legal obstacles and, thus, to the related moral deterioration of the workers. The development of military production, in this respect, increases the conflict among the competing firms because their desire for increased profits conflicts with limits on selling their products. Some known cases dealing with firms illegally selling equipment and technological innovations to hostile countries bear witness to this.

Moreover, a country's economic potential deteriorates because of a less effective management mechanism, which usually accompanies military production. Military production, in countries with market economies, leads to a different mechanism of operation than that of peaceful production. This mechanism directs the efforts of the top managers of the firms and the engineering staff not toward the growth of effective production, but in exactly the opposite direction. In Western countries the methods of operation of military production differ in many aspects from the methods of the operation in a peaceful economy. A peaceful economy practically functions within the framework of a market mechanism whereas military production is based upon relations between the government, acting for the military, and the private producers of military goods.

In signing government contracts with firms for the delivery of military products the government must recognize the expenses incurred by the firms in order to organize military production. Thus the firms become interested in showing exaggerated costs of production, and this involves both the necessary investments as well as current costs of production. The government compensates firms for the costs they incur through higher prices and/or subsidies. Moreover, the U.S. Defense Department is allowed by law to help firms involved in the delivery of military products to acquire their equipment, build necessary buildings, and receive guarantee loans from private banks. For example, almost half the capital of Universal Airspace, which primarily makes military products, is owned by the government, which gave the firm land, constructed buildings for it, and provided production equipment (p. 231).

The consequences of this may be quite unfortunate and have a radically negative effect on the ability of a country to convert its production from military to peaceful. In fact, under these operating conditions, management and engineering staff involved in military production become accustomed not only to ignoring costs, but even to trying to increase them. In peaceful production, working in the free market, exactly the opposite is demanded from these workers — the ability to decrease the cost of production.

Melman gives an interesting example dealing with Universal Airspace (UA) showing how firms specializing primarily in the production of military goods experience difficulties with diversification. All attempts aimed at peaceful production failed because neither the management nor the engineering staff had any experience in the construction, production, or marketing of products in a free market. In the end, the top managers concluded that if there is any possibility for the firm to produce goods for civilian applications, an attempt must be made to sell the production rights to the appropriate firms.

In exceptional cases, UA may diversify via a full separation of peaceful production from military production, that is, the creation, under the wing of UA, of a completely autonomous firms, with different names, locations, and management. This separation of peaceful and military production is supposed to prevent these autonomous firms from acquiring the ideas and traditions of UA (pp. 232–33).

Another aspect of the relationship between the government and the producers of military goods is the latter's attitude toward the size of orders received. Firms, usually, try to acquire large orders from the government to increase their profit. Therefore, they may even resort to corruption. This is easier to accomplish because the bottom line for products made by the firms is not cost and because the criteria for judging their activity under a monopsony are much more ambiguous than for firms operating in the market.

Therefore, the development of military production in peacetime may decrease the economic potential of a country. Whether this is compensated by the growth in the country's power, economic potential being a part of it, is a separate issue. However it is clear that the decrease in economic potential below a certain level initiates a decline in the military potential of a country as well.

Because of the negative effect of military production on a country's potential, it becomes necessary to curb growth of military power once it exceeds the critical level as well as provide ample incentives for converting military capacities to peacetime output. The convertibility of military to peaceful production in the United States causes some difficult problems which have to do more with designing programs to accommodate the transition than with creating the potential (predisposition) for this. Relatively low level of military production and the existing pluralistic democratic mechanism, are, in principle, sufficient to start developing programs for convertibility.

This process is based on the following assumptions. First, it is important to overcome significant difficulties caused by the natural resistance of persons at the macro, as well as at the micro level.

At the macro level, this resistance stems from people who think that the current level of military production must be preserved, or even increased, either because they are afraid that its decrease may have a negative effect on the nation's economic activity or because they consider it necessary to increase defense. Regardless of the reason, these people do not adequately account for the opposite effects this increase may have on a country's potential.

At the micro level, the resistance stems from workers fearing the loss of their jobs. Consequently, it is very important to create conditions to overcome this resistance. First, the areas of use of the freed resources must be clearly and unambiguously defined. Second, the workers in military production should not bear the full weight of economic and social difficulties caused by a transformation to peaceful production. It is crucial to undertake appropriate programs for financing peaceful businesses, for example, retraining the workers.

The areas of use of the freed resources and the programs for planning them should be clearly defined. Melman distinguishes the following four types of programs (p. 191). First, there are the complex programs at the national level. The author considers several such programs conducted by government offices. They include increasing spending on education, urban development, and improving the environment. These expenses, in sum, equal, or are even greater than, the possible decrease in spending for military needs (p. 197).

Second, there is a program to improve the living situation of deprived families, who in the United States numbered in several millions. This entails providing these families with professional training, making it easier for them to finance education, improve their living conditions, and relocate.

Third, there are programs of economic assistance to develop single large cities. As an example, the author notes the needs of Detroit, other large cities in the state of Michigan, and New York.

Fourth, there are programs that reveal the possibilities of replacing spending on particular military projects with peaceful projects. The author gives a long list of such projects (pp. 200–2). For example, the rejected plan for cleaning the Great Lakes, for the sum of $141 million, is equal to the demands for the creation of one new command post connected with the possibility of a nuclear air attack; 20 college scholarships totalling $41,000 equal the cost of one B-52 mission in Southeast Asia.

Additionally, at the micro level, the programs for converting military to peaceful production must give the firms sufficient information about possible new products and how any inconveniences for its workers will be compensated. The author gives a list of military businesses in electronics, aircraft construction, and shipbuilding industries that may produce new products (p. 230).

Thus, the current political systems in the countries of the free world are the starting points in their potential to provide, in the combinational style, for a possible conversion of military to peaceful production, particularly if the interplay of different political forces begins to gravitate toward less military spending.

While I agree with many of Melman's critical comments on the negative effects of military production on the U.S. economy and society, it seems to me that there are noticeable gaps in this work. On the one hand, the author undervalues the necessity of significant military spending by the United States in the face of the danger from the then aggressive USSR, which was truly a military economy. On the other hand, the author exaggerates the negative effects of military production on the U.S. economy; this is even clear from the book's subtitle, The Dawn of American Capitalism.

NOTES

1. It is also possible to separate the expenditures from the point of view of the capacity to inflict strikes at different times. The expenditure for inflicting the first strike is defined by the expenses on the maintenance of the appropriate sectors of the military. There may also be expenses for conducting a second strike — up to the expenditures for conducting a long-term war. From this point of view, the direct expenses for developing the newest types of weapons, if the development of these weapons is still remote, would be classified as expenses for a far-off strike.

2. At the end of the 1980s during the period of Gorbachev's perestroika, the Soviet leaders acknowledged the aggressive nature of Soviet military strategy and stated their desire to shift to a defensive strategy. At this point, it is impossible to verify their claim.

3. I say "at best" because an exhausting industrialization of the country in the 1930s was not completely justified. The USSR was able to win the horrible war with Germany with approximately one-third of the capacities created before the war.

12

Macroeconomy IV: Mechanisms of Operation of Economic Systems

STAGES OF AN ECONOMIC PROCESS

Innovations and their final implementation in mass production represent a multistage process subdivided into at least six stages: basic science, applied science, design of new models, industrial implementation of the new models, design of new objects within the framework of the already implemented types and their industrial implementation (modernization), and mass (frequently repeated) production. This classification is based on the performance of methods, possibly different in each of these stages. The variety of methods is in turn stipulated by the combination between achievements in the managerial practices and the culture of the given country. As a result, different countries may be better adapted to perform different stages. Table 12.1 illustrates comparative effectiveness of these stages in the United States and Japan (Gomory & Schmitt, 1988).

It is reasonable to assume that the managerial system in Japanese firms has an edge when it comes to modernizations and their subsequent introduction on a mass scale whereas the United States succeeds at the initial stages of creation and implementation of

TABLE 12.1
Stages of R&D and Manufacturing

Stages	*Country with Edge*
1 — Basic science	USA
2 — Applied science	USA
3 — Design of new samples	USA
4 — Serial production	?
5 — Sample modification	Japan
6 — Mass production	Japan

innovations. The success of the Japanese derives its impetus from the effective system of implementation of new technological ideas elaborated primarily in the United States. Japan itself has not produced much basic pioneering research. This is indirectly confirmed by the fact that even those Japanese scientists who have won the Nobel Prize spent a great deal of time working in the United States. The United States was ill-prepared to compete with Japan in many sectors of the manufacturing branch. However, the leading role of the United States in the sphere of pioneering R&D has, as of yet, not been seriously challenged.

Perhaps these relative advantages and disadvantages stem from the culture of the two countries. The United States, dominated by Protestant ethics and the resulting notion of society at the service of the individual and lacking feudal traditions, creates more opportunities for developing science because the individual is not stifled by incontestable authority (in the realm of science as well) that represents collective interests. Conditions in Japan, with its highly elaborate Shintoist ideology and the cult of nature, that is, something tested, with its prevailing doctrine of individual at the service of society (which is in accord with the Buddhist concept), and with its historical tradition of blood/relation feudal ties were more conducive to devising effective ways of implementing ideas within the framework of a firm as one large family (and in the country as one extended family where the government cares about the needs of the people, unlike in the United States where the government is an evil). With everything that has been said, the problem of competition between the United States and Japan could be resolved in a number of ways. But first, I present a brief digression.

Perhaps it is advisable to distinguish two forms of the Ricardian principle — sequential and parallel. By the parallel form I mean exchange between countries specializing in areas where they have an advantage in terms of producing various goods, such as natural resources, materials, equipment, consumer goods, and weapons. Sequential specialization denotes division of labor along the aforementioned stages reflecting the process of creation and production beginning with science and ending with mass production.

While the parallel Ricardian principle encompasses primarily private goods, the sequential one incorporates exchange between countries involving a large number of public goods, such as scientific/technological developments (excluding patents). At the present time, if countries produce public goods that can be used by other countries, the latter receive them almost for free. It seems to be the country's own business whether it wants to produce public goods and make them accessible to other countries receiving, at best, noneconomic benefits. Introduction of the sequential Ricardian principle calls not only for noneconomic benefits to be derived from such exchange but also for resources to acquire private goods.

Present relations between the United States, a major producer of public goods for the world, and Japan, a major producer of private goods, is sufficient to illustrate the problems that may arise in this connection. We could chart two alternative paths of development for the United States and Japan. In one each country would specialize in an area in which it has the advantage, and in the other each country would create an environment conducive to the successful implementation of all the stages. The actual solution will probably be a compromise between these two extremes.

It seems to me that a plausible solution for the United States lies along the following lines. Assuming the United States is superior to Japan in developing scientific/technological pioneering ideas and inferior in their implementation stages, the United States should concentrate on the initial stages. But here the United States would lag behind the Japanese at the latter stages. Because these stages also differ in the methods of performance, the United States should perhaps focus on those resembling the initial stages and bearing a direct economic effect. Naturally, economic benefits derived from these stages should be considerable. The stage of implementing new models seems to fulfill these requirements.

In other words, innovative ideas should not be limited to R&D but should be supplemented by their implementation at a stage of serial production. This would generate a tremendous effect both from the sale of the new products as well as their technology.

Naturally, these innovative technologies would be far from perfect (not the cheapest possible), which could be achieved only at the stage of new sample modification. Also output quality would not be so high as it is in a stabilized mass production of the modified version. Nevertheless, the greater benefits of the new product, both for the consumer and the producer (for instance, cost reduction as compared with expenditures on the previous product used to fulfill the same function) should compensate for all these deficiencies. By the time these innovations are modified and introduced on a mass scale, other innovations should appear on the market.

Thus the U.S. competitive edge with respect to Japan could be improved by emphasizing strong points, such as science and the development of innovations accompanied by their actual implementation, but only at a stage of serial production. Possible modifications of these designs and their industrial implementation as well as mass production could be largely carried out by Japan (and in the final stages even by South Korea and Taiwan).

Here I do not mean to diminish the need to improve U.S. productivity at other stages, both those where the United States is flourishing (especially in basic science) and where it lags behind mass production of modified models. I only want to emphasize one of these stages where improvements are especially vital. To succeed at this stage, U.S. corporations need to undergo major restructuring because they are largely oriented to mass-scale production and somewhat inadequate in reaping the benefits of rapid modifications of novel technologies.

HORIZONTAL MECHANISMS UNDER THE CONDITIONS OF INDETERMINATION

Economic systems as a whole, including the spheres of production and perfecting, can function on the basis of two types of mechanisms, vertical and horizontal. In vertical mechanisms one participant has the right to impose a course of action on another participant. Correspondingly, in horizontal mechanisms participants have equal rights. With all the ambiguity of these definitions, they correspond sufficiently to our intuition used in distinguishing between a government's centralized control over the economy and a market economy. Naturally, every economic system has both vertical and horizontal mechanisms. Distinction in their proportions leads to phase transitions, to transitions from one economic system into another.

Indeterministic processes are realized primarily within the bounds of horizontal mechanisms. And precisely because a developing system holds perfecting and its realization as priorities, it makes horizontal mechanisms dominant and gives vertical mechanisms a subordinate role.[1]

As long as the indeterministic approach presumes that decisions taken by individuals must be subjective, one of the fundamental conclusions is that economic systems must be based on independent participants. These participants must be competent and responsible for their actions. Thus, the significance of the independent source of their income, including their savings, that is, decentralized ownership in all of its numerous variations. The more indeterministic the system (the less clear the connections between actions of a given participant and those of other participants), the more independent must be the source of income (and the more decentralized the ownership) and to the maximum extent possible belong to the individual.

Thus, for instance, an artist's creating new art must be maximally independent from the consumer. This first of all presupposes his access to income received independently of the results of his artistic activity. Such income may come from family money, which was especially characteristic of societies endowed with aristocracies. There are famous examples of talented artists and scientists, for instance, Paul Cezanne and Henry Cavendish, who could engage in creative activities because they had access to funds amassed by relatives. If a society is sufficiently wealthy and an individual's income gives him the means of self-support even if he works an incomplete work day, then one of the alternatives for the artist's independence includes his receiving income from work whose fruits can be valued by the society. Moreover, such work can give an artist opportunities to gather impressions necessary for his creative activity. To a great extent this also relates to men of science or to such among them without a need for expensive equipment.

Thus, the experience in the United States is of great interest. At the end of the 1970s, Alexander Riasanovsky, professor of history at the University of Pennsylvania, organized a lecture cycle at the Wharton School of Business dedicated to three great Americans: philosopher Alexander Johnson (1786–1867), composer Charles Ives (1874–1954), and poet Wallace Stevens (1879–1955).[2] They were all financially independent, combining their creative activities in arts and sciences with work in insurance companies. With all that, their business activities were rather creative. During Stevens's visit to the University of Pennsylvania, a student asked him whether working at the insurance company detracted from his writing poetry. Stevens replied that, to the contrary, his work for the company helped him in his artistic activity.

I know two gifted U.S. artists who gained independence from the consumer by working part-time for institutions whose activity combined well with their creative temperaments: one of them taught mathematics, and the other restored old paintings. I asked the artist who taught mathematics if teaching distracted him from art. He responded that, to the contrary, it helped because interaction with students provided him with creative impulses.

Obviously, the mentioned financial sources are insufficient to guarantee creative independence, especially in science, hence the importance of tenured positions in Western universities. They guarantee scientists the opportunity to select the area of their research and to continue this research even if their colleagues are not convinced of the significance of their intermediate results. This constitutes the cardinal distinction between universities and academic institutes where the scientist's work must be approved.

In the case that an artist cannot earn enough money and does not have family money or a tenure, he can receive help from other individuals or benefactors but not from organizations. This requirement has a rather serious basis because a pioneer cannot give clear justification for the significance of his results and a program for the achievement of even the most roughly formulated results. A benefactor appraises subjectively the pioneer to whom he extends support. He values primarily the pioneer's potential and not the result and its program of achievement. I stress the role of individuals in the organization of aid to pioneers because pioneering activities cannot expect help from a group even if it is private, for example, a private foundation or a firm. Groups generally demand collective decision making on the principle of full agreement among all or among the majority, and the majority's opinions, as a rule, conservative. To make the participants' opinions sufficiently objective, one needs to make the appraisal of expected results more objective.

If the creator wishes to sell his results, then the society must have individuals who can purchase them. To do this, these individuals must have sufficient income or amassed wealth. Only an individual can value a creative pioneering result, by giving it an aesthetic appraisal and by giving an aesthetic appraisal to the creator. If a group of people, even if private, needs to purchase something, such as a museum purchasing new paintings, this will lead to the difficulties mentioned above regarding the involvement of an organization in financing artistic activities.

Experience in the USSR can be used as a rather extreme illustration. Because there the arts were controlled by the government and all museums were national, the latter were allowed to purchase only paintings made in the spirit of socialist realism. At the same time the USSR had a small group of people with high incomes, for example, academicians and chief designers of military equipment. Among them are people willing to support new directions in the arts. Thus, in the 1970s academician Sergey Lebedev, one of the creators of Soviet digital technology, supported the young artist Vladimir Muraveiv by granting him the opportunity to live for two years in his apartment after graduating from art school. At the same time, Lebedev helped Muraveiv sell his paintings to his friends (as far as I remember several paintings were bought by academician Victor Ambartsumian). In the beginning of the 1970s, I visited the Moscow apartment of Oleg Antonov, one of the leading designers of airplanes. It had several paintings by the officially unrecognized Soviet modernists.

The next step toward financing artists and promising them certain independence is provided by foundations. On one hand, foundations can limit an artist to a greater extent than private benefactors because they demand a confirmation of their experts' suggestions. Meanwhile, experts can make the very worst appraisers of the new, if only because they are too used to known paradigms. Of course, if we talk about the appraisal of work done within the limits of known paradigms, the experts' role is invaluable because it can prevent a useless expenditure of enormous funds. On the other and, because there are usually a number of foundations and the artist is required to confirm only the significance of suggestions given for work, not its results, as happens on the open market, foundations

can play a role in certain stages. They can be useful both at the stage of fundamental science and when the proposed work is already in the stage accessible to experts.

Thus, the above considerations regarding horizontal mechanisms relate to the nonmarket class of mechanisms. Other stages correspond to the market class of mechanisms where the appraisal of final results is possible to a greater extent.

FLEXIBLE VERTICAL MECHANISMS UNDER THE CONDITIONS OF DETERMINATION

Generally speaking, vertical mechanisms largely correspond to a deterministic situation. If the goal of development is clear and one can create a program tying this goal with the realized condition, a centralized or even planned economic mechanism may be preferable. This condition arises in a war, during which even countries with market economies adopt centralized planned approaches. The creation of inflexible centralized mechanisms of planned control over the economy in the Soviet Union was conditioned precisely upon the military posture of Soviet society.

I believe that the Marxist concept of an ideal economic system is based on the most primitive version of such a deterministic system. For Marxists, the future economy is similar to Robinson Crusoe's household, or to a textile plant with clear objectives and resources dedicated directly toward fulfilling these objectives, foregoing all possible interferences from prices and money. This direct distribution does not preclude the existence of hierarchical control in the case of a large system. The formation of such a hierarchy relates only to the technical difficulties of communication between the center and the individual, similar to those in distributing newspapers to subscribers. The publisher cannot easily deal with each subscriber. However, in the case of powerful radio stations and radios, such a hierarchy may appear unnecessary.

Thus, the extreme opinion about the necessity of centralized distribution of resources has a basis only within a set of highly limited assumptions, as, for instance, in the case of a textile plant with fixed demand and fixed technology.

Under the general case of a deterministic approach, with possible degrees of freedom to choose the composition of production and production methods, selection becomes necessary. This, however, is impossible to achieve directly, without a pricing mechanism. For a long time it was held, especially by critics of socialism, that this requirement cannot be realized within the framework of vertical mechanisms and that, therefore, centralized control of a large economic system is impossible. Meanwhile, scientific research of this century, beginning with the works of Abba Lerner, Oscar Lange, Fred Taylor, and ending with decomposition algorithms for the solution of mathematical programming problems, showed that introduction of a pricing mechanism is also possible within the framework of a deterministic vertical economic scheme. In this case, such a mechanism becomes the important means for the effective programmed organization of behavior in a large economic system.[3]

Thus, a deterministic approach to the organization of an economic system within the framework of vertical mechanisms draws attention to the concentration of resources in one organization (this organization may exist on the level of a firm or its part, as well as in a ministry). What is the sense of wasting resources among several organizations producing the same product?[4]

Meanwhile, critics of this deterministic approach see the dire consequences of such a concentration of effort: absence of competition and the accompanying heightened responsibility for growth of the organization's efficiency. The situation becomes more complicated when we turn to the area of perfecting in terms of creating new technologies. Partially or, rather, to a great extent in these situations parallel organizations may appear (see Chapter 3).

However, within the framework of vertical mechanisms, the parallelism of organizations becomes too much ordered; that is, the level of indeterministic process exceeds the level of its organization on the part of the controlling body. This leads to rather serious complications. The nature of control in such a hierarchy is usually multilevel, in the sense that each given level of hierarchy controls several lower levels. Under these conditions, creation of a new organization, if it is called for by the given level of hierarchy, demands interference on the part of several higher levels. Moreover, the head of the new organization may be appointed by the higher levels or, at least, approved from a number of candidates nominated by the given level of hierarchy. Thus, the emergence of an independent organization leader from the given level is disallowed. We must take into account that in appointing leaders of subordinate levels of organization, the higher levels of hierarchy are concerned not only with these leaders' business abilities, but also, perhaps to a greater extent, with their loyalty toward their bosses. Because of this, the level of this type of leaders' creative professional talent tends to fluctuate cyclically. Within each cycle, this level drops until a crisis forces its elevation. The above applies not only to appointed supervisors, but also to the approval of nominees. Therefore, although the given organizational level creates the list of candidates, it takes into account the possibility of their confirmation by higher levels.

Thus, if the economic mechanism were limited to the optimal distribution of resources, I would have ventured that contemporary means of mathematical economic analysis and enormous computing capabilities could have allowed to create vertical mechanisms of control at least equal to or even better than horizontal or market. However, the economy includes areas of perfection and implementation of results. This is one of the breaks in the chain determined by the very nature of the perfection process, which necessitates horizontal mechanisms.

VERTICAL MECHANISMS UNDER THE CONDITIONS OF INDETERMINATION

The above in no way indicates that the state may play no part in indeterministic processes. But how it does so is very important, especially as it concerns formation of organizations targeted toward pioneering work.

Any interference on the part of the state bureaucratic machine in the formation of innovative organizations naturally creates the tendency in innovative organizations to function in a deterministic manner. This is due primarily to the necessity for control of the taker by the giver and the desire of the giver to forestall the taker's ability to put the received funds to less than optimal uses. Naturally, the degree of such interference on the giver's part may vary. This is why the understanding on the part of the giving organization of the degree of determination of the taker organization's behavior is extremely important. Different interactions form between the organizations, correspondingly. Insufficient understanding of this circumstance leads to the isolation of interactions between giver and taker organizations independently of the degree of the latter's indeterministic activity. However, in all cases, the tendency of the giver to control the taker remains.

The U.S. government's role in the formation of the space program and in the fight against cancer serves as an example. From the late 1950s to the early 1960s, the United States responded to the Soviet challenge by creating the space program directed by NASA. This program's success was overwhelming. As contemporaries and living witnesses to these events, we cannot in full appreciate the greatness of this organization's achievement. We are talking about no less than the creation of a new civilization. After the first steps were taken on the moon, we entered a new civilization, characterized by the possibility of human control over space.

Why were NASA's achievements so great? One of the fundamental reasons is that NASA's creation was based on a relatively low level of indetermination of this organization's processes. Each project had not only a clearly defined end goal, for example, to land man on the moon, but also the central means for the achievement of this goal, such as rockets. The problem was to create the program of action that would allow the realization of these means. This program constitutes a rather branched-out net, combining all possible parallel and consecutive processes whose completion was necessary for the achievement of the set goal. This net's branches were given to certain organizations with defined roles. The crucial difference of this net was in clear formulation on the organizational level of local goals and in the probabilistic definition of times to completion. These comments do not mean that solutions to all local problems in the net were completely clear, that is, that the organization's internal activities were wholly determined. The creation of means toward the achievement of some intermediate goals, such as the design of heatproof rocket insulation, may to a great extent be represented as an indeterministic process.

The analysis of NASA's projects in terms of the indeterministic concepts here proposed demands special investigation. However, on first approach, all we can say is that when the means toward achieving a defined goal are known and it is possible to create a program for its achievement, direct interference on the part of the state in this process, creating a state agency, may prove effective.

The government's interference in developing projects dealing with indeterministic situations presents more difficulties. As an example we can use the government project begun under President Richard Nixon for the fight against cancer. As was the case with NASA, the state issued enormous sums to this project and formed a head organization in the National Institutes of Health to distribute the funds. However, practically everything resembled a traditional system in which a foundation distributes money to grant proposals by individual scientists and institutions working in the given field. I do not wish to disparage the usefulness of such a structure of resource distribution and the effectiveness of expended funds. It simply seems to me that if the government wants to interfere in an indeterministic process, there are more complex methods available.

Let me begin by saying that foundations, as all organizations, are successful primarily in respect to established directions because within such a framework it is possible through a system of experts to evaluate grant proposals objectively. Meanwhile, under the conditions of indeterminism, the means toward achieving the given goal are unclear. By comparing the space projects with the cancer project, we can clearly see the difference between the deterministic and indeterministic approaches, of course, under the simplified, or binary, usage of these categories. In the first case, the end goal was clearly defined, to put a man on the moon and to bring him back, and the direct methods of its achievement, the creation of special rockets, were understood. In the second case, only the end goal is certain, to prevent cancer cells from surviving, growing, and spreading, if they threaten the body, and, moreover, to restore the normal functioning of afflicted organs. The means to its achievement are completely unclear. The latter circumstance stems from the absence of a concept explaining the nature of cancer to a sufficient degree to begin the creation of means toward eradicating it.

Thus, under indeterminism, the means toward achieving a goal are unclear, even if the end goal is known. Hence, the cardinal problem of organizing research in the indeterministic area is forming concepts that could lead to the development of such means.

As we know from the history of the space program, developing such concepts demands a radical change in the thinking process. I will not go into the general problems of social organizations that foster this type of person. From general observations, these must be pluralistic societies where a person can form his own world view and be independent. Thus, for instance, one of the pioneers of space exploration, outstanding Russian scientist Konstantin Tsiolkovsky (1857–1935) "caught" the idea of the necessity for space flights from the Russian thinker, Nikolai Fedorov (1828–1903), who developed the ideas of general resurrection of all dead and of preventing death through modern science. Stimulated by these ideas, Tsiolkovsky created the revolutionary concept of a rocket engine and defined the general working principles for such an engine as the means for delivering man into space where the residual "dust" from the dead may be found.

If one realizes that Tsiolkovsky worked in the beginning of the creation of motor-driven aviation, it is possible to imagine how his ideas were met by the scientific and engineering communities. Fortunately, Tsiolkovsky possessed a certain independence, because he taught mathematics and physics in a secondary school in Kaluga. And fortunately for Tsiolkovsky, he found enthusiasts willing to continue his work during his lifetime. Among them was the distinguished Soviet engineer Sergey Korolev (1906–1965), the creator of the first Soviet rocket that delivered a person into space.

As for cancer, because no concepts provide clear ways toward creating means to eliminate it, one of the primary goals in this area should be developing such concepts. It is possible to assume that this type of concept will come from sufficiently independent people, most likely, scientists. However, I imagine that with the help of directed financing and, possibly, with help from the government, this type of concept could be forced.

Because of cancer's effect on the functions of the entire body, it may be proposed that formulation of cancer concepts may demand the efforts of scientists from many different professions. Meanwhile, scientific work throughout the world is organized primarily by disciplines. This is why it is so important to aid the formation of interdisciplinary groups in the process of cancer concept formulation. Given this goal, an organization providing financing for cancer research could select a group of universities and help them to finance these types of groups. The groups may be formed only by the universities and scientists themselves. The organization of these groups must be flexible and provide for informal discussions of problems on the part of creative people. To do this, conditions for increasing the frequency of meeting among scientists must be created. Their offices and laboratories must be in close proximity to each other. This, in turn, necessitates giving this group a separate floor with a dining room serving free lunches and dinners.

NOTES

1. It occurs to me that analogous effects take place outside the socioeconomic systems. The brain, which is primarily oriented toward perfecting, guarantees, it seems, equality of the left and the right hemispheres: they do not have a coordinating center.

2. Johnson wrote other books besides the original philosophical study, for example, a book on economics (1813).

3. In fact, this also relates to the organization of Robinson Crusoe's household, if we suppose that he used rational methods in deciding what and how to produce (Katsenelinboigen, 1972).

4. This is manifested in Soviet conditions, according to Anatoly Fedoseev (1976): "As concerns civil facilities, then all efforts are made to prevent 'extras' and to make production standard in all facilities" (pp. 231–32).

13

Branch: Evaluation of the Potential in Strategic Mineral Resources

The concept of potential can be used to analyze any of various sectors in an economy. It is probably best applied, however, where the concern of policy makers goes beyond whether, on a short-term basis, supply, demand, and prices are within acceptable bounds. Many sectors of the economy — those involving defense and international trade, for instance — display additional important aspects that standard economic analysis describes poorly. One such sector is nonfuel minerals, and one leading analyst says explicitly that security assessments, apart from any economic analysis, are needed to comprehend the various political and institutional dimensions of minerals. I intend here to outline a holistic assessment employing the concept of potential to this sector of the economy. The illustrative examples concern the Soviet and U.S. potential in strategic minerals.[1] This chapter is based on part of the "Report to the National Council for Soviet and East European Research" concerning the development of a methodology to assess Soviet economic potential; this part was written mainly by Kenneth Myers.

A BRIEF SURVEY OF LITERATURE ON SUPPLY OF STRATEGIC MINERALS

For the most part, the literature dealing with the problem of strategic mineral supply is based on the class of programming methods, which includes both reactive and selective ones.

Illustrating reactive methods are procedures that seek solutions via extrapolation. Noteworthy in this respect are works by Wilfred Malenbaum (1977). He has prepared projections of demand of strategic minerals for the world and various regions. His central finding is that countries consume large and expanding amounts of minerals per unit of GNP in early development, but in more mature economies the marginal amount per unit of GNP decreases, even though total tonnage may increase. This lower "intensity of use" holds for all minerals except aluminum, which is used more intensely in advanced economies. The conclusion for the former Soviet Union was that, because demand is leveling off, pressure on production will not grow markedly. As demand in East Europe and friendly Third World countries rises, the Soviets could continue to meet it with modest increases in imports. All long-range projections show that the ratio of reserves to production in the Soviet Union may decrease, but depletion is not expected in most cases.

An example of the use of selective methods in this field is works that formulate the problem in terms of the model of equilibrium. Prominent in this area are words by Herbert Levine (1981) and his colleagues. As far as aggregate indicators of mineral production are concerned, these authors want to determine the minerals future output on the basis of a general equilibrium model, called SOVMOD (where mineral resources constitute one of the branches) of the Soviet economy (Green & Higgins, 1977). Regarding particular kinds of mineral resources (for example, chromium, cobalt, and platinum), the authors make use of the regular method of extrapolation. This method utilizes the results obtained by Wilfred Malenbaum along with some general conclusions related to the growth of national income obtained by the authors themselves from SOVMOD.

Levine, in particular, found that, because of expected slow or negative growth, there will be little pressure on mineral supplies, and, hence, the current capacity will be adequate for medium-term needs.

Some work undertaken by U.S. scholars in the field of strategic minerals has some bearing on the method of potentiating. In it, the role of relational parameters, especially as they concern the external conditions, has been made explicitly clear.

Generally speaking, the significance of external conditions can be analyzed in two different ways: the first one concerns the necessity of using external sources of mineral supply for domestic needs; the second one concerns superpower competition in the international market. Among relational parameters being considered in these works, self-sufficiency (closeness) of the country's economy also is mentioned.

Daniel Papp (1981) discusses current accounts and how the Soviets program efficient decisions within the constraints set by policy. Because costs are counting in an already costly industry within the lagging Soviet economy and low-cost imports are readily available, he claims that a long-standing Soviet policy of self-sufficiency is even now being abandoned. This is a questionable conclusion on two counts. First, no Soviet policy forbids importing, temporarily or as a habit, as long as it is not too extensive or drains too much hard currency. Second, the Soviets have never let high costs keep them from producing materials domestically if there is a military or political justification. Of course, there is a limit to this, but there is no evidence that the limit has been reached.

Translating these remarks into formal terminology, we might say that Papp has made too rigid and simple the model constraints contributed from the policy-making stage (self-sufficiency) and then magnified the feedback from the programming stage (the price difference between domestic and imported minerals) in order to force a relaxation of the constraint. The market analysis that Papp offers is often mistakenly thought to be complete. The mistake is possible because, for countries highly sensitive to market efficiency, such as the United States, much is explained. Also, this analysis bases itself on the large amounts of numerical data that are generated as a result of trading, and this is somehow reassuring. One could incorrectly assume that, if Papp were able to submit complete and reliable data, this same kind of analysis would adequately describe the Soviet system.

Admiral Mott (*The Resource War and the U.S. Business Community: The Case for a Council on Economics and National Security*, 1981), for instance, sees the Soviets as nearly invulnerable because they are nearly self-sufficient and suggests the United States should imitate them.

Simon Strauss (1979), an industry representative, asserts self-sufficiency as a leading indicator, but adds a more differentiated account. He shows how costs do not deter the Soviets. For instance, Soviet aluminum production (which supplies all COMECON countries) uses about nine tons of low-grade domestic ore to produce one ton of metal. Western producers use four tons of ore per ton of metal. The Soviets also keep consumer demand low so that no net imports are needed. (Soviet citizens use only about one-fourth the aluminum that Americans use.) An important aim is to export minerals to gain hard currency. A long-range, subsidiary policy is to influence or disrupt southern Africa in a way that raises the prices of the minerals that the Soviets also produce (gold, platinum group, chromium, manganese, and asbestos). And although they profess to want to aid and buy from Third World producers, this has proven to be a sham, as it would interfere with self-sufficiency. Finally, land use or pollution problems are not allowed to block new development.

Additions and refinements to this picture, sometimes contradictory ones, are offered by William Severin and Daniel Fine. Fine (1980) claims that the Soviets are beginning a "resource war" against the West. He feels that the Soviets are importing more and exporting less, partly because their own supplies have become scarce and expensive, but in their adversity they see political opportunities. If they can both assure foreign supplies for themselves in the long run and disrupt Western supplies, they reap a double benefit. For instance, efforts to increase state control over mining are encouraged in African countries. It is rumored that Soviets (without needing the supplies) bought large amounts of cobalt just before the mines were shut in Zaire by guerrillas. This helped the price climb, and some Western aircraft manufacturing was interrupted. The Soviets, according to Fine, are seeking secure overseas supplies for the day they will depend on them and in the meantime will take advantage of opportunities that arise in the course of trading.

The Bureau of Mines Soviet expert, in a ghost-written article (*Mining Annual Review*, 1981), takes a dim view of the resource war idea and offers some evidence that he believes contradicts it. Soviet exports have high priority in the national plan. It was never the case that the Soviets exported surpluses that could not be used internally; hence, increased internal demand would not automatically cause a reduction in exports. In many cases, exports are even sold below cost, which supports Strauss's findings. A possible reason for reduced exports is the price rise for some of them (petroleum, gold, platinum, and diamonds). The quota for foreign currency earnings can be met with less tonnage in these items, and there is less pressure to export the other "surplus" minerals that are now needed much more at home. These are manganese, asbestos, titanium, vanadium, and nickel. Soviet chromium, in addition, has become less valuable in world markets because of changes in technology, hence, lowering the incentive to export. Furthermore, although the Soviets import some materials, only with tin and certain rolled steels are they net importers, as they pass along much of the material they import to COMECON countries. The weakening position in lead and zinc, with sharp new imports, can be linked to the cutoff of lead and zinc concentrates from Iran.

Severin (1981) also finds that the resource war idea is a little far-fetched. Although there is importing, he thinks it will be temporary because large new mines are coming on-line and because the Soviets will continue to use mining as a way of developing Siberia and the Far East. Because international mineral trading will never become great, he does not find that the Soviets will have enough market influence to seriously damage the West. He says that self-sufficiency was never exactly or completely the policy; the aim is rather a secure and large production that develops remote areas and that keeps domestic and East European industry well supplied. There was never any prohibition of overseas buying.

In summary, these are the main themes discussed in the Soviet mineral situation. Despite production problems involving cold weather operations, construction bottlenecks, lowering quality of ore, and capital shortages, the traditional level of Russia's self-sufficiency will not change dramatically. Regardless of whether preventing any imports is the main determinant of policy, the tradition of developing the outer areas with new mining could continue; exports could also continue, although reduced in tonnage as receipts have risen, allowing a greater level of consumption of mineral production at home. In many cases the mineral is produced above world prices.

VULNERABILITY AS THE EQUIVALENT OF "LACK OF POTENTIAL"

Parameters of Potential

A system behaves well in a changing environment if it avoids specialization. This means that it needs to retain favorable strengths, abilities, and relationships to the environment, even if these are currently unused or insignificant. Table 13.1 lists 27 basic parameters of strengths and abilities. They were based on the list of primary relational parameters mentioned in Chapter 8.

While revisions and new categories are possible, this list has proven to be a useful and comprehensive guide in identifying the key factors to a particular system's predisposition for development. Because policy makers intuitively understand potential as a criterion, one should expect to find it used in some partial way along with a list of relational parameters. This beginning effort can then be improved because the principles behind it have been generalized.

Several organizations developed a national strategic minerals vulnerability index. Among these organizations can be distinguished the Strategic Studies Institute of the U.S. Army War College (King, 1981). On the basis of these materials, one can clearly see that in the nonfuel minerals literature "vulnerability" is clearly meant as the equivalent of "lack of potential." The most obvious evidence is Chapter 15 in the book by Bohdan Szuprowicz (1981). It shows how one would summarize the various factors that make a nation or corporation vulnerable to mineral supply problems. Of his 26 vulnerability factors, only one did not easily correspond to a parameter of potential mentioned in Chapter 8. Ten parameters of potential initially had no match.

Szuprowicz explained what principle he used to formulate the factors. He relegated to a separate section a discussion of specific threatening scenarios and estimates of their likelihood. In his mind, the vulnerability factors stand alone as an index to the strength of a supply system, under an assumption that the mineral environment is essentially unpredictable. Using my terminology, he separates potentiating from programming exactly according to my definitions by not assuming a predictable environment in one section and then making assumptions in another that facilitates the discussion of specific sequences of events. He makes it clear that the assumptions of the policy-making stage more accurately match the true situation.

Kenneth Myers extended the search for vulnerability factors and found 61 additional factors, plus many repetitions. He reconciled these with the list of general parameters of potential (as they have been mentioned in Chapter 8). In some cases the list was found deficient and was changed, but in other cases a parameter suggested a topic that was very poorly covered in the literature. With additional searching, sources were found that agreed that these neglected topics were, in fact important. Thus the general list of parameters was not forcing irrelevant distinctions. This reconciliation process is an interpretive task that cannot be done mechanically, and there were many minor adjustments, which cannot be described here. The result is a corrected list of parameters consistent with a policy-making perspective (in other words, distinct from models of current accounts and from mere statements of belief about the system, yet it can be related to each of these efforts), within that perspective, that states all the major concerns expressed in the minerals policy literature and includes an interpretation of all the general parameters of potential.

Although these corrected parameters are intended to be exhaustive, they are intended to overlap somewhat and, hence, are not exclusive. This is because holistic judgment, which characterizes policy making, is not reducible to exclusive operational definitions. If such definitions were presented, they would misdirect an expert away from using his experienced judgment and toward calculating quantities, which may or may not be good indicators of the general parameters under consideration. The only way to tell whether they are adequate indicators is to quiz experts for their judgment, and we are back to the need for a list of factors that match the generality and overlap the ideas found in the policy literature.

These factors lend themselves to the following use and interpretation, which can be seen as a procedure for completing some of the decision-making functions. A country's priorities in improving its vulnerability, or its views on what are the controlling variables, can be expressed as weights assigned to the vulnerability factors. These weights are contributed from the valuing stage. Myers offers the weights that would probably be assigned by Soviet and U.S. policy experts for their respective countries.

Second, a particular mineral's vulnerability can be rated with reference to these factors. This offers a structured way to identify system changes, but also a holistic way to evaluate and prioritize various threats. Using the terms introduced previously, current decisions can be evaluated with respect to the criterion of mineral vulnerability in order to build a mineral position with high potential in the face of the unknown future. I do not claim that this is a new activity, only that the concept of potential transforms a common sense technique for measuring vulnerability into one that is self-conscious and formally developed.

TABLE 13.1
Vulnerability Scores

Soviet	*U.S.*	*Item*
Location Factors		
M	M	1. Closeness of sources. Both ship over long distances.
L	L	2. Dispersion of mines.
L	M	3. Route protection. Soviets have natural protection. U.S. transport could break down or be attacked in many places.
Political and Market Factors		
L	M	4. Monopolization. Soviets, monopolists themselves, promote centralization. U.S. watches for, but does not fear, damaging cartel action.
L	M	5. Suppliers independence from trade. The Soviets import large amounts only from dependent suppliers such as Guinea and Cuba, expressing a strong sensitivity here. U.S. wants to assume that everyone must trade, but recognizes exceptions, such as Iran.
L	M	6. Trade and investment restrictions. Soviets welcome host government control. U.S. depends on relatively unregulated commerce.
L	M	7. Ideological compatibility. Despite rhetoric, Soviets have no compunctions over trade with enemies, and few reject them. The U.S. is rejected more often and is pressured to embargo, for example, Rhodesia, South Africa, and Soviets.
L	H	8. Supplier instability. U.S. preoccupied with stability problems, thinking they may lose favor with new regimes. Soviets apparently unconcerned in the few places they trade.
L	M	9. Access during crisis. Soviets depend on domestic reserves in crisis. U.S. uncertain over ability to extend force to reopen supply lines.
Domestic Strength Factors		
H	L	10. Mineral industry size. Soviets are determined to keep a large industry. U.S. has been content to mine less.
H	H	11. Domestic reserves and resources.
M	H	12. Stockpiles and scrap. It is not clear whether the Soviets stockpile extensively, but they probably have less interest in it than the U.S.
H	M	13. Domestic processing capability. Soviets want complete technology. U.S. content to have many processing stages performed overseas.
M	M	14. Equipment upkeep. Soviets continue to build but have poor maintenance and chronically run under capacity. U.S. is not replacing old plants, except in aluminum.
Domestic Flexibility Factors		
M	H	15. Substitution potential. A U.S. preoccupation, possible panacea. Topic seldom mentioned with reference to Soviets.
H	M	16. Reallocation potential. The Soviet command economy is based on this concern. The U.S. places less emphasis on this but has plenty of capability nevertheless.

TABLE 13.1, continued

Soviet	*U.S.*	*Item*
H	M	17. Recycling/recovery. Soviets have recently raised the priority on curbing waste, which is rampant in the minerals industry.
L	L	18. Use in combination.
M	M	19. R&D in use and processing. Soviet priority was low here, but technology now recognized as badly dated.
L	H	20. Regulatory costs. Soviets have many bureaucratic and domestic policy constraints but seem willing to bear this. Both the U.S. producers and government feel this is a key to reviving the industry.
M	H	21. Business costs. The Soviets had a lower priority here but raised it as costs could no longer be ignored. The U.S. government has tried to ease costs through tax breaks but is very reluctant to offer any subsidies to producers, no matter what the mineral.
M	M	22. Plant expansion lag. Soviets were low here, but decided to concentrate on finishing projects before beginning new ones. U.S. lags are long now because of regulation and investor emphasis on quick returns, raising government concern.
H	M	23. Surge capacity. The Soviets always have mobilization production in mind. The U.S. seems to be correcting its neglect here.

Weights

The weights in Table 13.1 express which factors policy makers may feel are generally most important and which are controlling variables in the assessment of vulnerability. The weights reflect beliefs, beliefs that may change on the basis of experience. The description of these values is based on both the stated intentions of a country (what it wants its actions to mean to others) and what habitual actions appear to mean at this general level.

The weights as linguistic variables are H (high), M (medium), and L (low), meaning that an item tends to be of high or low importance when compared to all the other factors. The weights and comments will, in effect, be a redescription of the Soviet policy and strategic situation as indicated above.

Group Weights

The rank order of the groups of factors is the following (highest priority = 1):

Soviet	*U.S.*	
3	4	Location Factors
4	1	Political and Market Factors
2	2	Domestic Strength Factors
1	3	Domestic Flexibility Factors

The Soviets were relatively unexposed to political and location problems because so much was mined and processed domestically. Their concerns were in their ability to manage and adapt their assets. The United States feels comparatively confident about its long-run flexibility, despite problems, and is resigned to having to ship from far-off regions. In fact, some believe that, given this initial exposure, the best way to deal with it is to expand trade as broadly as possible to create new options and alliances. With any policy, the political and market factors are a high priority in controlling vulnerability. These judgments on the group weights will be used to adjust the item weights somewhat.

Interpretation of Weights

Adding the weights, we have the following:

Weights	*Soviets*	*U.S.*
L	9	3
M	8	4
H	6	6
Total	23	13

Because the United States has higher weights overall, this implies that the United States will tend to have higher vulnerability scores for any mineral. The justification for this difference is twofold. Because the United States is exposed with many minerals and because they interact, an extra measure of caution should be added. In addition, these weights reflect a nation's ambitions and self-perceptions, and there is no objective stance from which the weights would be in identical proportions. One might object to the limitations this causes in a comparative procedure. It remains comparative in that the same process is applied to both, which makes explicit the differing decisions and values that shape the economy of each country.

The parameters of potential must now be rated by the criterion of potential or vulnerability. In other words, are options and abilities retained that might be useful in any of a variety of situations? All the minerals could be considered together, but that would be perhaps even more of a strain than was assigning weights.

Scoring Procedure

Although the ratings can be interpreted in their current form, they can also be numerically manipulated in the following way:

Weights:
for groups rated 1st and 2d H = 9, M = 5, L = 2
for groups rated 3rd and 4th: H = 8, M = 4, L = 1

Vulnerability scores on each mineral item: H = 3, M = 2, L = 1

A high vulnerability score means that, compared to other minerals, this mineral is in a more vulnerable condition. The weights and the scores are multiplied and will determine whether the exposure is serious.

The minerals can also be judged on whether the factors are probably changing for better or worse, which would help indicate where preventive action is needed. Table 13.2 presents the scores combined into summary form.

Discussion

The variety of results offers many points worth discussing. A few of the more interesting questions are taken up here. They do not press the analysis beyond the preliminary evidence.

Soviet titanium is rated nearly as low as is possible in this technique, and the other metals are not especially high, judged against the U.S. cobalt score, which would be among the highest of any U.S. scores. A few Soviet minerals, such as molybdenum, would give them a high score. That the Soviet aluminum score is above the U.S. score is no cause for concern for the Soviets — small differences at a lower level are meaningless in this rough rating scheme. The basic message, however, is that the Soviets are not particularly vulnerable, but neither are they without any worries. The United States, also, has a few major minerals in which it is very well situated, despite the fact that a good portion of them are imported. This rating scheme has forced a variegated comparison of the two countries.

According to the literature, the Soviets had short-term supply problems. These are important, but they do not appear to have corroded any fundamental strengths as indicated in these ratings. The supply problems have, rather, convinced the Soviets to raise the priority on shortening expansion lags, reducing costs, and upgrading technology so that unfavorable trends in these items get extra attention, and this change has been reflected in the weights.

Looking at the weighted scores, we see the United States is vulnerable in cobalt and has an adequate position in the other two. Other minerals, such as chromium, manganese, or possibly platinum, would produce higher scores. These are somewhat less interesting cases because they offer no opportunity to produce domestically as an option to increase importing. With cobalt and titanium the

TABLE 13.2
Weights, Scores, and Trends for Aluminum, Cobalt, and Titanium in the Soviet Union and the United States

	Soviet			*United States*		
	Weighted	*Unweighted*	*Trend*	*Weighted*	*Unweighted*	*Trend*
Aluminum	179	(37)	4 down / 0 up	166	(31)	2 down / 1 up
Cobalt	206	(42)	1 down / 0 up	255	(51)	7 down / 0 up
Titanium	135	(27)	3 down / 1 up	167	(33)	3 down / 1 up

United States has expressed its choice to import most supplies whereas the Soviets chose to subsidize their domestic production. The difference can be explained by a higher level of values, which is only indirectly expressed through the weighting. This would fall within a discussion of the valuing stage of decision making, but one might say that the United States has a higher tolerance for vulnerability in general, expecting that its flexibility factors can take it through any adverse changes (reflecting the ideology of laissez-faire Yankee ingenuity). This was not the U.S. opinion following World War II or the Korean War, when shortages were fresh in mind, but these things are easily forgotten. A second point is that the United States, even more than putting a high priority on cost factors, believes that the country is definitely aided as a result of consuming only lowest cost minerals. This expands the decision-making cycle by taking into account national potential, where a savings in one system allows the country to expand resources in other high priority areas. For instance, subsidizing cobalt production was rejected recently because of efforts to reduce the federal budget, which is expected to help the whole economy regain vigor.

The unweighted scores are shown to get an additional perspective on how the different sets of weights affect the scores. This would be helpful to show experts as they are asked to revise their weights. The change in proportion between raw and weighted scores shows that cobalt gets its exposure a little less than does aluminum from the sensitive or high weighted factors.

In general, a high score, weighted or unweighted, can be a policy warning. A country may want to attend to one factor of one mineral, a factor across all minerals, or all factors of one mineral. The decision makers have to decide whether high scores at any of these levels are controllable and worth the effort to change. A decision to reduce vulnerability would involve, minimally, a change in programming constraints.

The trend judgments can be very helpful in formulating these decisions. If potential is deteriorating badly, the sources can be pinpointed, but rather than simply trying to reverse that trend, many more options are readily apparent. For instance, favorable trends may counterbalance the unfavorable ones in the overall score and, thus, argue for no action. Also, more effective countermeasures may be found by changing another factor, which can be moved, than by fighting an item in inexorable decline. Some people are worried, for example, that most U.S. chromium smelters have shut down, which increases factor 13, domestic processing capacity, and less so factors 1, 2, and 23. The industry has changed such that no amount of tinkering, short of large-scale subsidies, would revive the smelters. The shipment of the low-grade ore now used is costly, and, at many large new mine sites, smelters are being built. By buying from the

new low-cost sources, favorable changes could be effected in factors 2, 4, 5, 8, 9, 20, and 21, and perhaps overall vulnerability would be improved. Of course, policy makers would have to recheck all the judgments they made that contributed to such a result. If they really did not believe that abandoning domestic smelters improved vulnerability and if they were interested in consistency and in taking all factors into account, they would have to change a few weights and scores and find some holistic justification for the decision.

The discussion and comparison of minerals can become much more detailed, but this presentation has been sufficient to show how potential provides a framework that can keep a multifaceted problem in perspective while parts of it are analyzed, and it does this without intruding confusing technicalities. Most important, it does not often contradict the careful opinion of experts, whose judgment is mirrored in the results, yet the results are more differentiated than an expert can usually muster in unaided discussion. Also, factual data can be used to argue for changes in the judgments because the judgments are not at such a high level of generality that they are simply guesses or unreasoned opinions that are immune to further discussion.

NOTE

1. Although the USSR is no more, much of what has been said holds true, if not for all its former republics, at least for the leading one, Russia. I believe Russia will find a way to maintain economic ties with former Soviet republics and possibly with some new ones which may form as a result of Russia itself splitting up. Let me add that the Soviet Union was a very unreliable source of information; and, to the extent the Soviets trade minerals, trading partners contribute incomplete accounts. Nevertheless, there was much correct and timely information scattered in Soviet publications. If information about specific plants and mines is assiduously collected, as it is at the U.S. Bureau of Mines, a relatively accurate general picture can be pieced together.

14

Firm I: Potentiating at the Firm Level

The classical theory of the firm expounds at least three basic postulates. The first of these states that the firm's interests and those of its head concur. The second is that these directly revolve around profit maximization. The third postulate states that at each moment of time the decisions by the firm's leaders can be completely linked to profit.

Harvey Leibenstein (1979) called attention to the fact that physics and biology in recent decades have made tremendous progress because they have developed by expanding the boundaries of their research. This expansion has been accomplished primarily thanks to the analysis of microstructures. At the same time, his opinion is that in economics, generally speaking, no such sharply defined trend toward pushing out the boundaries of research is visible. Twentieth-century economic theory, after being revolutionized by Keynesianism, has mainly occupied itself with the mathematical interpretation of previously defined economic ideas. Leibenstein believes a whole branch of economics has been lost, one he calls micro-microeconomics. This branch ought by rights to concern itself with the problems entailed by abandoning the aforementioned postulates of classical firm theory. According to Leibenstein, the importance of these questions is so great that they may hold the promise of enabling economics to overcome its backwardness in expanding the borders of its research.

Given all the interdependencies between these three postulates of classical economics, I would prefer to examine them separately, utilizing in this a conception of the external and internal functions (external and internal criterion) for the activity of an economic unit.[1] This chapter's principal attention will be turned to revision of the third postulate.

INTERESTS OF FIRM LEADERS VERSUS INTERESTS OF THE FIRM ITSELF

Examination of the interconnection between the firm leader's interests and those of the firm itself has been conducted in the economic literature on both a starkly theoretical and an applied level. Russell Ackoff and Fred Emery (1972) presented the problem in its most general form in terms of internal and external functions in the interrelations of people and organizations. This is what they have to say:

> If an individual or system displays a type of structural behavior that is a member of a functional class of behavior displayed by other entities, then that individual or system can be said to have an *extrinsic* function. It is called extrinsic because the function is not one of its own but one it has by virtue of its membership in a class. . . . The individual displays an *intrinsic* function because its function can be attributed to it on the basis of *its behavior alone* (p. 26).

Analogous considerations about external and internal functions can be applied to a firm and its leader. A firm's drive to maximize profits could be thought of as its external function. It is characteristic for all firms regardless of structure and is dictated by the general drive of all firms to output the most valuable production and economize the most valuable resources. At the same time the firm's leader drive to obtain more of fewer rewards from the firm can be interpreted as its internal function.

Victor Polterovich (1968) casts the problem of the interrelation of external and internal functions in a more applied way, in terms of the external and internal criteria of an arbitrary economic unit. This author's view in somewhat simplified form may be presented as follows. The unit's decision-making organ, represented by a person, forms on the basis of his individual values, an internal criterion, X. The value of this criterion depends on the stimulus vector z, which includes all the things the individual receives, such as material rewards and moral support. A feedback mechanism, in turn, incorporates the external optimality criterion for the given economic unit, p (profit, say), and a stimulus operator, R. The latter determines the stimulus vector, z, depending on the value of the external criterion.

If the firm's leader is also inseparable as its owner, its external criterion, all else being equal, unambiguously defines the value of the leader's internal criterion. Thus, if profit is the firm's external criterion and the sole component of its owner's stimulus vector, profit singularly determines the value of the owner's individual (internal) function. If in this the firm's leader increases personal expenditures, labeling them as business expenses (or reduces them), this is equivalent to an increase (or decrease) in personal income; the firm's income and the leader's personal income correspond. (Naturally, I am neglecting here all the possible tax complications under which the method of expenditure attribution, business or personal, makes a difference.) The situation changes if the firm's leader has partners (or shareholders) or is simply a manager whose income is a part of the firm's expenses.

A number of business researchers, and in particular Tibor Scitovsky (1943) and Oliver Williamson (1963), have turned their attention to the point that a manager strives first of all to increase personal income in the broadest sense of the word, including therein not only monetary recompense, but also a luxurious office and the best secretarial services.

In this instance a well-known conflict between the firm's external criterion and its manager's internal one arises, that is, between the manager's private income and that remaining to the firm. Of course, one could object to this opposition of expenditures on managers and firm interests by arguing that establishing the best conditions for management enhances productivity and, thus, helps the firm's interests overall. But one way or another there is a problem in defining the degree of these expenditures. In his book Lee Iacocca (1984) describes in great detail the conflict over this issue in the Ford Motor Company during the rule of Henry Ford II. The latter expended enormous sums of money on his own personal needs (for example, buying expensive wines in Europe), charging them to the firm.

Thus, in this situation businesses encounter the complicated task of coordinating external and internal decision-making criteria and constructing a stimulus vector for managers that provides the best possible coordination of their interests with those of the company.

My subsequent attention shall be directed mainly toward analyzing the firm's external criterion under the assumption that it is operated either by its owner or an automaton, devoid of personal interests.

Methodologically, the definition of a firm's external decision-making criterion may best be understood if we imagine it as part of the whole mechanism of the economic system. Such a methodology corresponds to a systems approach that demands the system under analysis first of all be placed within a suprasystem. The links between this system and its suprasystem enable us to better comprehend the demands made on the former. It is then possible to analyze the system, that is, its internal mechanism.

According to classical theory, a firm's external optimality criterion is to maximize profits as the difference between expenditures on the acquisition of resources and revenues from the sale of production. The realization of such a criterion assumes at least the following two conditions: knowledge of the prices of resources purchased and of products sold and opportunity, in fulfilling the preceding condition, to coordinate completely and consistently all resource expenditures with a specific sales return.

WHY PROFIT IS INSUFFICIENT TO EVALUATE A FIRM'S PERFORMANCE

In general, every firm has degrees of freedom that permit it at any given moment to choose what resources to procure and what to produce bearing in mind its internal capabilities. In order to realize these degrees of freedom, the firm must have information from its environment about the available options for supplying itself with these resources and the potential sales of its output. If the environment contains an enormous number of firms, then, generally speaking, the given firm cannot directly possess such global information about its environment because to do so would require it to know the internal production possibilities and resource requirements of all other firms (as well as households and the government). Therefore, in choosing a course of action the firm needs the necessary global information about its environment in a condensed form.

Research into the various operating mechanisms of economic systems, as well as hypothetical scenarios for new mechanisms, shows that such global information in condensed form can be represented by using different kinds of parameters. In particular, the conditional and complete prices of output produced and resources consumed may serve as global condensed-form parameters to guide a firm's behavior.[2] The notion of conditional (complete) prices was introduced in Chapter 8. Unless otherwise noted, the term "price" will signify complete price. Under certain conditions, detailed in works on mathematical economics, prices may be a sufficient condition for firm self-determination. They direct its development toward producing more valuable output and economizing the most valuable resources.

Thus, the existence of prices gives us the chance, under the proper conditions, to form a holistic evaluation of a firm's behavior. This evaluation may be given through its reduction both into individual products and their prices and into individual resources and their prices. In other words, both the output and expenditures can be directly represented in the firm's weight function in the form of a linear polynomial, that is, as the sum of the values of output produced and expenditures incurred. The difference between these — profit — also serves as the basis for measuring the total efficiency of the firm's behavior.

All this about the role of profit in evaluating a firm's activity also remains true if price information is probabilistic. Here, knowledge of prices in terms of probabilities presupposes that some mechanism links all the participants of the economic system in a consistent manner. But profit is an insufficient criterion for evaluating a firm's performance.

Depending on whether the causes of this are rooted in the firm's external connections with its environment or within the firm itself, they may be considered internal or external. If we assume isomorphism between the processes of establishing equilibrium between the firm and its environment and the processes taking place within a complexly organized company with a large number of internal links, both the external and the internal factors conditioning the combination of profit with other indicators will likewise be isomorphic.

In the more general case, which shall also be the subject of our further analysis, the external factors that cause profit to be insufficient for evaluating a firm's performance are connected with the firm's lack of complete links with its surrounding environment. In this event it is impossible to construct sufficiently complete prices on which to base a profit criterion. Incomplete prices result — prices that partially reflect the firm's connection with its environment. The relational parameters and their evaluations that are added to the material parameters with their incomplete prices completely fill the breach left by the absence of the proper links. An analogous situation in the general case exists with respect to internal factors.

What I have just been saying about profit's role in organizing firm behavior applies entirely when a firm's production process takes place instantaneously or, more precisely, in the space of an irreducibly small interval of time. However, a firm's production processes are usually prolonged. Thus a large number of intermediate products emerge. Their production is important to the firm's development sequence. In principle this kind of assumption in no way conflicts with a profit maximization criterion. Under these conditions, if long-range prices exist, the search for a profit maximum as an integral over time is the key to determining the amount of intermediate goods to produce at the current moment.

With an approach that is differential over time, that is, taking place at each point in time, profit will be only a part of the evaluation of the firm's performance. In this case, along with profit, the internal links between the output of intermediate goods and the resources expended on them must also be considered in terms of the external links between the value of final output produced and resources expended for this. Here the situation is completely analogous to constructing a differential evaluation of the macroeconomy based on national income as an indicator. The latter, as it is known, includes the output of consumer goods and increment of capital investment. If there is no need to pay nature, the value of consumer goods produced becomes analogous to a firm's profit, and the increment of capital investment equals that of intermediate goods production.

Meanwhile, difficulties in establishing prices on the firm's intermediate goods that will take into account their value and utility for the future make it so much harder to provide the appropriate evaluation of the firm's performance at each time interval. Determining the value of intermediate goods on the basis of their production costs is usually insufficient because it does not include their future utility. The whole process of balancing a good's value with the cost of its production is missing.

In order to make it possible to establish appropriate prices for intermediate goods, methods for coordinating resource expenditures with final goods output have to be found that take into account the large number of intermediate stages. However, finding such methods is not always possible. There may be two reasons for this: either the managers are unable to perform the proper calculations, even though they theoretically have all the necessary information (they may be busy with other work or lack fast computing equipment), or the information is unavailable.

What are the various methods for solving the problem of coordinating profit with its complementary indicators if there is no way to tie together in a unified model all the final results and initial resources with the intermediate results? I shall try to examine these methods independent of the variety of external and internal factors that dictate the need for such coordination, but with a certain emphasis on the former.

To reiterate, in the previous sections I proceeded from the extreme assumption that prices were required to measure profit and that this indicator (profit) is sufficient to evaluate the firm's performance. Now, let us introduce a different extreme assumption regarding the definition of an efficiency criterion for the firm. Suppose that the firm does not know the prices with which profit can be measured. Such a situation may take place, for example, when the firm has to make decisions in a turbulent environment. Because the links between input-output parameters are unknown, price determination for future periods becomes impossible. Thus, the problem of evaluating a firm's performance is acutely complicated when the growth dynamic is explicit and the evaluation must be integral over time.

Here the problem of the planning horizon emerges. It may be defined as that period between the present and the future within the limits of which prices may be established that are in principle sufficient to measure firm efficiency using profit. If the firm is indifferent to what happens after reaching the horizon, there are no problems with evaluation. But how can a firm's post-horizon performance be evaluated if its condition toward the close of this period exercises an impact on its further behavior. Under these conditions, if the firm is merely optimizing its profit (or a mathematical expectation of profit) over the one period for which prices are known, it will not be well enough prepared to face the unknown beyond the confines of this period. This pertains to both final and intermediate goods.

The lack of clear links between the firm and its future environment makes it necessary to focus special attention on the firm's preparedness to face the future outside the bounds of the planning period in relation to both final goods and intermediate goods output.[3]

EVALUATION OF A FIRM'S POSITION

Methodological considerations expounded in Chapter 3 regarding the multiphase approach to the study of economic processes will facilitate our analysis of the problem of evaluating a firm's performance. Here, I make one strong assumption that the decisions are made by a single person; for now, multiplicity of participants in the decision-making mechanism is not considered.

Possibly it is useful to begin study of this problem from the general assumption that developed links between firms in the economic system are unformed and that the individual firm does not know what decisions to make concerning the structure of its future input-output. Random search procedures are well-adapted to this phase. As the links between firms improve, the possibility increases of a first shift making it possible to measure the firm's condition and characterize the degree of its predisposition to further growth under conditions of an unknown future. Once the next phase is reached, it will become possible to develop a mechanism that will allow for elaborating many competing strategies of development, for the subsequent selection of the alternatives addressing the current needs of the firm, and for monitoring the performance of the selected alternative and its replacement if necessary. In view of the fact that each of the constructed strategies cannot be corroborated unequivocally, they all must be included as part of a confederate-type mechanism; the picked strategy must conflict as little as possible with other strategies.

Once the strategy reaches a stage when its creator can evaluate the state of the firm, the firm is in a position to undergo the transition of the second phase. What distinguishes this phase is the feasibility of measuring the predisposition of a given strategy toward further development. Here, the firm's managerial structure can assume more federalistic shape.

As the firm evolves and interacts with its environment, it may reach a state amenable to the transition of the third phase. This marks the firm's transition into the environment, which allows probabilistic prices of various goods and resources to be set, prices that are, at least in principle, sufficient to fulfill the role of global parameters that integrate the economy for the duration of the period ranging from the current moment to the set horizon. The traditional criterion of profit, the difference between costs and revenues presented in probabilistic terms, becomes an extreme case of the preceding phase at which only the predispositions to future development are established.

As development conditions for the firm and its environment become more certain, a point is reached where a last shift is possible. This shift is characterized by the chance to derive with full certainty prices sufficient to function as global parameters integrating the economy in the period from the present moment up to the planning horizon. Traditional profit criterion, such as the difference between cost and revenues, becomes with these prices an extreme case of the general problem of measuring firm efficiency.

The more general case from which one can begin to study firm behavior is the situation where a multiphased spectrum of performance evaluation methods can be assumed. This spectrum has two extreme states: one when global parameters do not exist as prices and the other when they do and with full certainty. Between these two extremes are states in which firm performance can be evaluated. Although global parameters do not exist as prices, there is enough information to give a crude answer to the question of the firm's preparedness to face the unknown future.

What is it that characterizes these states? Theoretically, there also arises the problem of measuring the value of firm performance in relation to them as well. As soon as the firm expends limited resources and produces goods, the difference between cost and benefits is, like profit, subject to evaluation. However, we cannot think of this kind of profit as the difference between the sum of the values of the individual products and individual resources. In this case the firm's evaluation may be given merely as an integrated whole without the chance to differentiate the values of individual products and resources.

How is this integration accomplished? Can it be represented analytically or would that conflict with the need for wholeness? In other words, can an assemblage of integrative parameters be found that will enable the firm to be correlated with its environment? What is the meaning in this case of the evaluation of the individual goods and resources that take part in the exchanges between firms? Only in the extreme case, when interfirm links are so developed that profit can be introduced as the difference in the sums of the global values of individual products and resources, does it become possible to correlate directly the general firm evaluation with the evaluations of particular individual goods and resources.

I can be asserted from general considerations that the integral effect can be represented analytically as the value of the weighting function, if not only the material parameters that constitute it are separated out as independent variables but also the relations between these (primary parameters); by the same token the weighting function must take into account the evaluations (conjugate parameters) of these independent variables.

Thus, the primary material parameters might be production and sales volume of the appropriate goods, the amount of reserves of a resource or product, the equipment on hand, and so on in real terms. The conjugate parameters would be the unconditional prices on the corresponding material parameters. Examples of relations could be as varied as the firm's market share, the variety of output it produces (level of diversification), its level of vertical integration, and so forth. Their evaluations (also conjugate parameters) must reflect their influence on the firm's future development.

The difficulties of reducing the integral effect to its analytic representation is conditioned by the following factors: first, how does one define the primary material parameters; second, how does one measure their evaluation without having the complete conditions describing their interaction; third, how does one determine the large number of various relations between material parameters, and fourth, how does one evaluate them? It must be borne in mind through all this that the evaluation of individual ingredients and that of the relations between them have to be in comparable terms: otherwise the integral cannot be given a total evaluation. In other words, evaluating a firm's performance under conditions of an unknown future requires the synthesis of both the individual parts that make up material parameters and the relations that describe their organization.

The task of my further analysis will be to reveal the conditions under which it is possible to render such an analytical representation of a firm's integral development without sufficiently complete and available information about the firm's future functioning. Of course, at some point information will be so meager that it will be impossible to give a more or less reliable picture of a firm's effectiveness. But, I repeat, we are interested in the conditions under which it will, nevertheless, be possible to evaluate firm performance given incomplete information about the economic system's functioning conditions.

Let us begin by looking at how the value of individual ingredients may be established in such cases. These evaluations cannot be derived from any prior system state because prior states did not include the new conditions that might appear in the future. By the same token, future prices cannot be determined by extrapolating from past periods. Doing so might even be disorienting because actual future price correlations could be significantly altered by new conditions. In connection with this, the need arises to construct a new kind of prices, which could be called unconditional or incomplete prices. These prices must take account of certain ideal future development conditions, which the economy can asymptotically approach independent of any concrete conditions both with respect to participants' intentions and the initial resources they have.

I do not even know whether the problem of developing such prices for firms has even been raised. This issue requires research of its own. It seems to me that one potential way to solve it is to find the turnpike theorem prices mentioned in Chapter 9. Unconditional prices must be supplemented with relational parameters characteristic of a state in which the ongoing processes do yield to observation.

As for relations, they have yet to hold the central attention of researchers into firm behavior, with the possible exception of indicators like a firm's market share. At the same time it must be assumed that the role of relations in evaluating a firm's performance is far more significant than had been supposed. Interesting in this respect is the work of Hiroyuki Itami concerned with qualitative analysis of such "invisible assets" of the firm as consumer trust, control over the marketing of the product, and the culture of the corporation and its managers. Summoning the experience of Japanese companies, Itami made a strong case for the importance of these invisible assets in strategic planning (Itami, 1987). One piece of research that stands apart is the PIMS model. The main objective of this model is to express formally the quantitative weights of different parameters and above all the relations that influence the evaluation of firm performance.

Unconditional prices, combined with relational evaluations, will have to show the degree to which the firm is currently prepared to face the unknown future, how the environment's development can be induced in the desired direction, and how the firm is prepared to take advantage of beneficial unforeseen events while sweeping away undesirable ones.

The integrated evaluation of the horizon's position is sharply subjective. The reasons for this are very deep. Evaluation of the horizon's condition presumes that the firm's head will be able to realize the position created in post-horizon periods. Meanwhile, how things will look at that point is unknown. Also unknown is how the chief will act in that future situation. Therefore, in evaluating the horizon's position, the chief must take into account his potential to realize that position in the future. Thus, what may be highly valued by a leader of one caliber and style, might be evaluated completely differently by someone else. Thus, when we speak of subjectivity, we are at the same time combining several factors: the intuitiveness (nonformalizedness) in problem solving and specificity of the level and style of the executor (itself the result of the synthesis of his biology and culture, including professional).

The decisive role of subjective evaluations of the position by no means prevents each component of its analytical representation from having to be subjective. Constructing an evaluation of a firm's position requires an organic combination of objectivity and subjectivity. If unconditional prices may be formed objectively, based on information about technologies, relations become more complicated. We cannot tell whether a list of relations figuring into the evaluation of a state is exhaustive. New, quite complex, and profound relations could be found based on the experience of the particular individual. These relations might be hidden in intuition and cannot always be formalized. Chess, a system simpler than economics, is an example that seems to support this point. But even if one assumes a formalized list of relations, the evaluation of these is thoroughly subjective because it reflects not only experience that lends itself poorly to formalization, but also the unique level and management style of the particular individual.

This does not mean that in evaluation relations cannot be formally constructed. One can strive for freedom from the subjectivity of the specific leader. That kind of objectification and the accompanying formalization of the process can be achieved both on the basis of statistical analysis of the past and on expertise. The evaluations established in this way might prove quite useful. All this does not exclude or replace the subjectivity of a firm's head but is organically combined with it. The evaluations obtained in this formalized manner might be of great aid to a leader in decision making. They may provide a definition of the boundary conditions to stay within. They may (depending on their coherence) also stimulate the leader to explain why a subjective evaluation diverges or to provide confirmation that the latter is not just the fruit of deluded perceptions, as is not atypical of intuitive problem solving.

Let me now comment on the creation of a state oriented toward some distant future. Unlike chess where one and the same player conducts the entire game, in the real world the set planning horizon can exceed the professional life span of a given employee in the assigned position. In the distant future other executives are likely to manage the firm — something like a relay race. If new leaders are appointed based on how well they harmonize with former leaders, the established strategies will square with the future leaders. If the leadership is assumed by a person very different in his personal and professional qualities, it then makes sense to try to objectify the valuations of material and positional parameters to make these

valuations person nonspecific. In other words, averaged-out valuations are highly revealing when the features peculiar to singular objects are not known. In unique situations peculiarities may be so pervasive as to negate the worth of the generally recognized patterns.

Does this mean that the formation of positional parameters be postponed until a plan horizon — a position that will evolve into an indeterminate future — is reached? Not at all! It is vital to observe positional parameters before reaching the horizon, particularly for mid-term planning ranging from three to five years. There are many reasons to do so. One obvious reason is that the starting situation at the beginning of the period is not pregnant with too many combinational opportunities, opportunities to set clear-cut, narrow and program-attainable objectives aimed at maximizing profit. Here, it is necessary to continue to develop the position.

Another consideration affirming the need to track relational parameters is the system's inertia. It may be difficult, if not impossible, to attain in one quick dash the advantageous values of positional parameters upon reaching the plan horizon. This is the reason why the weights of these parameters ought to be modified over time. Here, the intricate problem of positional sacrifice comes to the fore, that is, sacrifice in profit for the sake of improving positional parameters. But as the economic affluence and the economic turbulence both grow, thus cutting down the time required to reach the plan horizon, problems engendered by a positional sacrifice (other conditions being equal) become a real issue.

PATHOLOGICAL POTENTIATING

Everything said about the necessity of the method of potentiating stemmed from the premise that managers of a firm cannot combine all the firm's indicators into a single indicator of profit or, for one reason or another, consider this inappropriate for themselves. This notion should be distinguished from the entirely different notion of the insufficiency of evaluating a firm's activity through profit under conditions when it is expressly impossible and primarily when the prices are spurious. That is precisely the sort of situation which took place in the USSR and in a number of socialist countries imitating the Soviet economic mechanism with respect to the evaluation of enterprises' activity.

Under the centralized planning system functioning there enterprises are assigned several dozen other indicators, along with profit, which plays a secondary role. These, in particular, include the gross output volume (including unfinished production), volume of sold production, nomenclature of production being manufactured (including the most important one, the so-called "red positions") in a physical expression, average salary of an employee, labor productivity, returns on capital, and unit cost of production. At the same time, the decisive role is played by the gross output volume. This sort of situation arose in the USSR in the 1930s during the formation of the centralized planning system, which completely ignored the accomplishments of Western economics and the latter's recognition of price as an economic invariant, that is, a category equally acceptable for both a market and a planned economy.

Moreover, the Soviet planning system ignored the conception of optimal planning offered as far back as 1939 by the outstanding Soviet scientist Leonid Kantorovich, who won the Nobel Prize for economics. This conception organically joins construction of the plan with forming prices and utilizing the criterion of profit to evaluate the activity of enterprises (as well as any other link of the hierarchical system). The price mechanism placed in the foundation of the Soviet planning system was viewed separately from planning in physical indicators; the conception of setting prices was founded on the Marxist labor theory of values, according to which in the plan, as in the center of gravity, the cost of the product must equal its production cost, that is, the average socially necessary expenses of labor during the manufacturing of the product.

Under this sort of price mechanism, a dualism arose in the system akin to schizophrenia, that is, actions guided by disjointed values (Katsenelinboigen, 1978). According to one set of values, based on the intuitive consideration of relative usefulness of benefits by the plan's constructors, the demands concerning output of production would be formed; the other set of values, based on costs, was used in measuring expenses, and prices of manufactured products would be set according to it.

Confronted with this sort of reality Soviet planners were forced to abandon the clarification of appropriateness of manufacturing corresponding products on the basis of prices and according to the profit evaluation of the enterprises' activity. In the early 1930s Stalin proclaimed the principle of "higher profitability" instead of profit for the evaluation of any economic measures, as well as the activity of any economic units. In practice this principle meant simply that practical workers should be guided by political considerations and common sense in forming the structure of manufactured production.[4] Because it was still impossible to ignore the necessity of summarizing indicators and comparison of general expenses and results, the indicator of the nomenclature of manufactured production in a physical expression was complemented by the system from the multitude of economic indicators.

The discussions intensively carried on in Soviet economics over the improvement of the economic planning mechanism within the framework of accepted Marxist economics were specifically concerned with decreasing the number of planned indicators of the enterprise and with ultimately returning to the economic methods of substantiating evaluations of activity on the basis of the single indicator of profit. That was precisely the essence of the economic reform launched in 1965. However, this reform quickly stalled, and, for the most part, the system returned to an expanded multitude of indicators characterizing the activity of the enterprise. The 1965 reform failed for political rather than economic motives. Its orientation to economic mechanisms of administration threatened the authority of the party apparatus, whose participation in the economy was justified by the existing administrative methods. But it is possible that the failure accomplished more than any success would have. In fact, it is fortunate that the reform did not succeed because introducing profit as the decisive indicator of enterprises' activity while retaining, in principle, the existing methods of planned price setting could led to serious breakdowns in the economic mechanism with all the corresponding consequences.

The projected transition to profit as the main indicator of enterprises' activity under the conditions of the perestroika begun by Gorbachev had, in principle, the most serious of foundations, for a market was supposed to be introduced with its corresponding price mechanism. It is another matter that this is incredibly difficult to accomplish, for a problem of unprecedented difficulty arose here — the problem of transition to a market economy of a huge military authoritarian-imperial industrial power currently in a state of crisis, with its human, natural, and industrial resources depleted and with its basic institutions, like absolute moral values and private property, destroyed. Because of the backwardness of Soviet economics, the role of the market and its classical institutions like prices and profit has been enveloped in a romantic haze and is extremely idealized. In my estimation, leading Soviet economists do not sufficiently comprehend the power and limitations of the market, above all, from the point of view of indeterministic development of the economic system under the conditions of rapid technological changes.

NOTES

1. Leibenstein's work looks at these postulates together. The single function that guides a manager takes into account profit, sales, rate of growth, and payments in kind (p. 479).
2. Interesting in this connection are mathematical methods of solving large-scale optimality problems. They include various diverse modes of interaction between the whole and its parts. Examples of such methods are the decompositional methods worked out by Dantzig and Wolf and by Kornai and Liptak.
3. One can apply here terms from optimal control theory — conditions of transversality.
4. I can recall a related case, which took place in the late 1960s, from Soviet life. Grigory Chukhrai was not just a very talented film director; he continually

strove to develop his talent. At the same time the organizational framework of Soviet cinematography and the constant arbitrary intervention of various organizations into the production of a film, from approving the script to releasing the film to the cinemas, greatly hindered the director's activity. At that point Chukhrai thought of creating a so-called cost-effective film company, which would shield him from the arbitrary rule of numerous bureaucrats. With this in mind, at a reception by the country's chief ideologue, Mikhail Suslov, Chukhrai asked him something like, "Distinctly formulate for us ahead of time all of the criteria by which we should be guided in producing films needed by the country." Then he added, "We commit ourselves to strictly follow these criteria. But after that allow us to do everything without any intervention from above." As far as I know, no reply ever followed.

15

Firm II: General Quantitative Methods of Evaluation of a Position

Our analysis so far was predicated upon the assumption that even if the integral (over an extended period of time) criterion of firm performance is profit maximization, its differential criterion can be modified because prices of the firm's final output and the prices of consumed resources do not convey complete information, which would interrelate the interests of the firm with the firm's environment. Incompleteness of prices is inherent in the operation of the economy under the conditions of indeterminism. Here, the differential criterion of profit is modified and assumes the form of the firm's potential. In other words, the firm's performance is evaluated by means of incomplete prices of material parameters posing as independent variables and the relational parameters evaluated accordingly.

REACTIVE MODE-HEURISTICS

I shall proceed to examine a number of cases where the constituent components of profit must be established individually because they cannot be interlinked directly. This task can be accomplished on the basis of the intuition of employees from the firm's respective subdivisions (or consultants) or in a formalized manner, for instance, by extrapolating the future values of individual parameters on the basis of past experience. Thus, the advertising department might determine advertising expenditures in future years based on their own intuitive knowledge or by formally constructing a function of advertising expenditures over time based on past experience.

The isolated search for an individual parameter value is based on all possible heuristics, the individual independent rules. Here is what Franco Modigliani (1958) had to say on the subject:

> In an oligopolistic situation, with its precarious internal equilibrium, there is much to be gained from simple and widely understood rules of thumb, which minimize the

> danger of behavior intended to be peaceful and cooperative being misunderstood as predatory or retaliatory. . . . The experience of those who, like myself, have conducted extensive personal interviews with executives suggests that these respondents have a strong propensity to explain their behavior in terms of simple mechanistic principles (p. 226).

One may attempt to separate heuristics into two groups: binary and continuous. Binary heuristics represent rules whose domain of application may be expressed in terms of "yes or no." The idea of using binary heuristics to analyze a firm's behavior was advanced in the work of Robert Gordon (1948). He calls these heuristics "shortcuts." These shortcuts take the form of commandments like: "Thou shalt not allow thy market share to go below 10 percent." "Honor thy dealers above all." "Thou shalt not unionize."

Richard Cyert and James March (1963) also speak about the large role of binary heuristics in the internal decision-making mechanism. They justify this by saying that precedent plays a major role in businesses because when the interests of different departments collide, institutionalizing past experience greatly eases the task of administration and saves the firm's leaders from having to resolve a large number of conflict situations. Here is what the authors of this work had to say about the role of heuristics:

> Past bargains become precedents for present situations; a budget becomes a precedent for future budgets; an allocation of functions becomes a precedent for future allocations. Through all the well-known mechanisms, the coalition agreements of today are institutionalized into semipermanent arrangements. A number of administrative aphorisms come to mind: an unfilled position disappears; see an empty office and fill it up; there is nothing temporary under the sun (p. 34).

Continuous heuristics pertain to situations when individual rules act as functions linking a limited number of parameters. For instance, expenditures to be allocated on advertisement are often estimated through these heuristics — by devising a rule to determine the amount of advertising expenditures as a percentage of the volume of sales.

William Baumol (1967) provides numerous examples of positive and negative uses of heuristics and makes highly interesting comments regarding the application of this technique to decision making. But the heuristic methods of measuring firm performance were elaborated most fully in the work of Peter Drucker (1954). This was indeed a pioneering work in management theory, for it questioned

the preeminence of profit as the sole criterion of firm performance. In his article (1991) Drucker mentions three mutually competing criteria of firm performance, all outside the immediate domain of profit. Following is the chronological sequence of the three criteria. The first one, advanced about 1950 by Ralf Gardiner, expressed the need for "the best-balanced interest" of the stakeholders. However, this criterion was hard to test if only because the meaning of "best" as opposed to "balanced" is unclear.

From that time to the present, the leaders of all major U.S. companies claim their primary objective is "to maximize shareholder value," that is, to achieve higher prices for a company's stock. This criterion is more concrete and realistic, but it suffers from its own Achilles heel. The trial period for the value of a company's stock is limited from half a year to one year whereas real tangible results from new capital investments might require many years. A policy aimed at a short-term boost of stock prices might result in highly detrimental consequences upon the company's future performance.

In following the tradition of the great British economist Alfred Marshall, Drucker proposed another criterion of firm performance, namely to "maximize the wealth-producing capacity of the enterprise."

> It is this objective that integrates short-term and long-term results and that ties the operational dimensions of business performance — market standing, innovation, productivity, and people and their development — with financial needs and financial results. It is also this objective on which all constituencies depend for the satisfaction of their expectations and objectives, whether shareholders, customers, or employees (p. 112).

Chapter 7 of Drucker's book (1954) is devoted specifically to an in-depth analysis of the indicators to be used in measuring the development of this area of a firm's operations. According to Drucker, the way to find the values of those indicators is primarily by means of heuristics. The heuristics are formed based on a set of indicators that diagnose eight major areas of a firm's operations.

First and foremost is the firm's market standing. The primary indicator associated with this area is the firm's market share by product, both new and established. This indicator does not omit the dynamics of sales growth relative to the actual or the potential behavior of the market.

The second area has to do with innovation of new products or technologies of production. In a number of manufacturing sectors, such as pharmaceutical and synthetic organic chemicals, the rule of thumb is that, for the company to survive the replacement cycle, its product line must be three-quarters complete in ten years.

The third area is productivity and contributed value. The latter basically refers to value added, a well-known indicator. To judge a firm's performance properly, it is vital to pursue an increasing ratio between the contributed value and total revenue, as well as the proportion of contributed value retained as profit.

The fourth area covers the supply of physical and financial resources required by the firm to pursue its respective goals for market standing and innovations.

The fifth area is profitability. It is commonly held that on average the return on capital should be about 25 percent before taxes.

The other three areas of a firm's operation include manager performance and development, worker performance and attitude, and public responsibility. It is more difficult to find indicators (not to speak of quantifying them) corresponding to these three areas. However, a number of quantitatively amenable indicators are available. Indicators for employee performance include employee turnover, absenteeism, safety, calls on the medical department, and grievances.

All these indicators of firm performance proposed by Drucker bear close affinity to the ones I proposed for evaluating the firm's potential. Perhaps, the only difference is my emphasis on characteristic parameters incorporated into the potential. Moreover, both Drucker's approach and my own share the notion of subjectivity of appropriate indicators and in asserting the need to find the magnitudes of these indicators for each firm individually.

Based on essentially independently set minimal value of the aforementioned indicators, Drucker attempts to delimit an adequate operating space for the firm. Developments in this area form the groundwork for subsequent evaluation of the firm's performance.

The principal feature different in Drucker's approach toward an expanded evaluation of firm performance is its focus on heuristics. All in all, as beneficial as heuristics are, their use is quite limited. Besides being incomplete and crude, they may also be inconsistent. The demands of one heuristic may conflict with those of others because, by definition, heuristics are independent rules, and this gives rise to the complex problem of their integration.[1]

A situation analogous to that with heuristics exists in chess. Chess players realize that a given heuristic is sometimes irrelevant and as a whole the heuristics are often mutually inconsistent. The analogy between heuristics in chess and firm behavior is quite cogently formulated in a brochure by Rudyard Istvan (1984) where he says:

> Rules of thumb are guiding principles that, while never true because oversimplified, point reliably to the probable direction of action. Experience builds rules by remembered results of trial and error.
>
> All good chess programs depend on rules of thumb to simplify calculations. Some moves are not examined, because they almost never pay. Very possibly, the brilliantly innovative winning move will be excluded by such a process. But the frustrations of novice players matched against the simplest chess programs demonstrate their power.
>
> Most of the major ideas in business strategy are conceptual rules of thumb about economic competition. The experience curve, the growth-share matrix, average costing, and the environments matrix are all rules of thumb. They help to pattern competitive behavior. They point toward the relevant and away from the extraneous, and suggest probable courses of action. They simplify, but cannot substitute for the thought process (p. 5).

I share the opinion of the authors of this brochure. Far from denying the value of heuristics, I think that, because of their inconsistency, the principal problem in charting a firm's strategy remains thoroughness in correlating all the diverse parameters used to evaluate firm performance.

Similar views are found in other works. Copeland, Koller, and Murrin (1990) note that as recently as 15–20 years ago company managers could perform effectively using the so-called "rules of thumb" (pp. 4, 7). However, the high turbulence of today's environment has made this approach less effective, calling instead for an overall valuation of the firm's performance. Here, the authors stick to a rather orthodox position by linking this valuation with the cash flow and its correlation with profitability.

Achieving greater congruity among material and relational parameters hinges upon new methods allowing for explicit comparison of the parameters' relative contributions (which ought to be marginal). I am not familiar with any works dealing specifically with the viable methods of computing such valuations. The works mentioned in Chapter 9 under the turnpike heading deal with highly aggregated parameters that preclude their practical implementation at a microlevel. Moreover, computational techniques covered in these works deal only with the methods of evaluating material parameters.

The PIMS model tackles the problem of parameter intercomparison, but its scope is limited primarily to relational parameters using the parameters' averaged-out valuations arrived at on the basis of a thorough examination of statistical data. Upon examining the PIMS model, the strengths and weaknesses of its methodology become apparent.

PIMS MODEL

The PIMS model — Profit Impact of Market Strategy — stands apart from other research in the area of evaluation of firm performance. My analysis of this concept is based primarily on the work of Sidney Schoeffler and his colleagues. In Schoeffler's (1980a) own words the project's objectives were:

> The intent of the program is to provide business and managers and planners with tools and data for answering questions like these:
> a) What profit rate is "normal" for a given business, considering its particular market, competitive position, technology, cost structure, etc.?
> b) If the business continues on its current track, what will its future operating results be?
> c) What strategic changes in the business have promise of improving these results?
> d) Given a specific contemplated future for the business, how will profitability or cash flow change, short-term and long-term? (p. 5)

In the course of their work, the authors of the PIMS model were faced with a number of basic questions regarding the analytical representation of a firm's position. They have managed to answer them, but within a pragmatic rather than a theoretical framework. Their approach was to use statistical data to develop the tools necessary to calculate the analytical weights of various indicators reflecting the firm's market position.

Initially the PIMS model was designed to accommodate the needs of a single company — General Electric. In the 1960s then vice-president Fred Borch, head of the marketing services division, initiated a program aimed at developing statistical methods to determine the numerical impact of various factors, relational for the most part, on company profits. G.E.'s uniform accounting system coupled with its immense experience in operating a highly diversified company (its line of products consisted of about 80 different items) provided the information necessary to design an effective cross-sectional model of the firm. Five years of intense research yielded a model consisting of a single rather complicated equation in which the level of profitability (return on investment — ROI) was correlated with such primary relational parameters as the degree of vertical integration, investment intensity, market share, and new product launches. The results obtained from the model were widely used by G.E., and it became a standard tool in company planning operations. Sidney Schoeffler supervised the work on the model.

From 1972 to 1974, while affiliated with the Harvard School of Business, Schoeffler utilized his experience at G.E. to start his own program, which in 1975 was converted into an independent institute called the Strategic Planning Institute (SPI). SPI developed a model based on the one used at G.E., this time incorporating the experience of about 200 companies including 125 from the Fortune 500 list. Altogether these companies accounted for more than 1,700 types of different products and services. The data base used for subsequent statistical analysis was formed from the information supplied by each firm in a special format, which allowed SPI to reduce it for the purposes of running the model. Naturally, special provisions were made to ensure the confidentiality of the information the companies supplied to SPI.

SPI conducts research based on this data and makes the results available to the companies for use in their planning operations. People utilizing this information range from the middle echelons of the company's managerial staff to its top executives. Their decisions encompass a wide range of strategic issues having a three- to five-year horizon (Schoeffler, 1980b).

What then is the methodology behind the model? First of all, the authors assumed that the ultimate goal of a corporation is to maximize its profit or, to be more precise, its profitability, the ratio of profit to investment — ROI). According to Bradley Gale (1986), the current version of PIMS incorporates many different criteria used to evaluate the firm's economic status, including D(CF + FMV), where D stands for discount value, CF for cash flow, and FMV for future market value. In my subsequent analysis of the PIMS model, I shall, for the sake of simplicity, use the profitability criterion for evaluating firm performance.

The PIMS model incorporates the very basic idea that a firm's current status together with current market conditions affects the firm's future prospects. In order to evaluate the firm's position on the market, or in the authors' own words, to determine the effect of different factors on firm profitability, they introduced a certain set of parameters into their model.

These parameters, from a qualitative point of view, could be as essential as relations. In the case of essentials, the authors have chosen only those whose direct influence on profit usually is very difficult to obtain, for example, advertisement and R&D projects. The relational parameters could be of any kind, structural or characteristic. The majority of parameters in the PIMS model are relational.

From a quantitative point of view, the PIMS model identifies an isolated subset of significant indicators having a 70 percent to 80 percent share in the overall evaluation of firm performance. Altogether there are about 30 different characteristics incorporated into the model. Below is an illustrative list of these characteristics:

Characteristics of the Business Environment — Long-run growth rate of the market, short-run growth rate of the market, rate of inflation of selling price level, number and size of customers, purchase frequency and magnitude

Competitive Position of the Business — Share of the served market, share relative to largest competitor, product quality relative to competitors, prices relative to competitors, pay scales relative to competitors, pattern of market segmentation, rate of new product introductions

Structure of the Production Process — Capital intensity (degree of automation, etc.), degree of vertical integration, capacity utilization, productivity of capital equipment, productivity of people, inventory levels

Discretionary Budget Allocations — R&D budgets, advertising and promotion budgets, sales force expenditures, strategic moves, patterns of change in the controllable elements above

Operating Results — Profitability results, cash flow results, growth results

The PIMS model is not limited to determining the impact of each individual factor on firm profitability. In other words, an equation developed by the authors is not a separable function of several variables correlating profitability with the values of individual arguments. It actually determines the cumulative effect of several variables taken simultaneously; it represents a nonseparable multivariable function.

For instance, in one of his works Schoeffler (1980c) presents statistical data on the interdependence between market share and the degree of vertical integration. Large market share happens to be especially effective when combined with high levels of vertical integration. Statistical data also indicates that market share is less significant if the business in question is more unionized than its competitors. Schoeffler (1980d) also showed that productivity growth is especially effective when achieved at low levels of investment intensity.

The advantages in the firm's position are potential in nature: the firm must also be able to realize them. This constitutes an extremely difficult and painful task. Sometimes, in order to improve the parameters affecting a firm's position (its future prospects), it becomes necessary to make sacrifices in current status. To handle this problem, Schoeffler (1980c) introduces a special term, "harvesting." It means a willingness to settle for lower profitability in the present in order to increase future market share.

The following question arises, "What specific steps should a firm take in order to realize its potential to increase profitability?" In this respect, Buzzel and Chussil (1983) have given a rather specific meaning to the concept of potential. Potential means the ability of a firm to take advantage of every available opportunity to improve its position on the market by utilizing the experience of other businesses having a similar strategic profile. Here, the authors distinguished between "sleepers" and "terminally ill" businesses. Sleepers are capable, at least in theory, of improving their position using patterns of behavior exercised by other businesses under similar circumstances whereas terminally ill businesses are not. The differences between these two kinds of businesses may be well concealed, making it difficult to distinguish between them.

Now I would like to say a few words about the philosophy, or general methodology, used by the authors of the PIMS model.[2] This model, which aims to quantify the effect of various factors on profitability, uses a methodology similar to that the natural sciences use. In fact, Schoeffler (1980a) states that one of the major objectives of the project is: "to discover the 'laws of the marketplace' that govern a) profit levels, b) other outcomes of strategic actions, and c) outcomes in the changes in the business environment" (pp. 5–6).

These laws of market and the laws of natural sciences have, in the author's opinion (1980b), the same nature. Just like the physical laws, the laws of the marketplace for Schoeffler (1980a)

> Indicate that the same principles that determined business success in the past will continue to operate in the future. All of science — physical, medical, or economic — is based on that assumption. (For example, if you expect the future market growth rate of one of your businesses to be 9%, materials prices to rise at 12%, and one of your competitors to drop out, the profit-impact of these events on your business will be determined by the same "laws" that operated in the past on other businesses in that situation.) (p. 19)

Starting with the model's methodology, there are things to be said in defense of the PIMS model. As we know, long-range forecasting based on attempts to establish some sort of pattern in the output and the pricing of individual products has not been very successful. Although econometric models of Western economies have been beneficial, we cannot deny the large discrepancies between the actual and the forecasted results for forecasts covering more than one or two years (Kolata, 1984). The limitedness of these models is attributed mainly to two factors: omission of a number of significant parameters and incomplete information on the ones already incorporated into the models.

It seems to me that, unlike traditional econometric models where the emphasis was placed on trying to forecast specific material parameters corresponding to input-output components and their respective prices (sometimes even expressed in aggregate terms), the PIMS project focused primarily on establishing predispositions based on the interplay among various essentials and relations and their impact on profitability independent of the specific nature of the product. On the basis of general considerations, it seems that the latter model is better equipped to handle long-range forecasts because it incorporates relational parameters and their effect on the final outcome as well as their mutual interdependence. These parameters are more stable than direct links between various input-output components (even if they are presented in aggregate terms) because the latter are subject to frequent changes.

At the same time, the authors of PIMS seem to be overstating when they claim to have discovered laws of the market analogous to those in the natural sciences. The authors of PIMS make a serious attempt to handle the difficulties associated with trying to establish the laws of the marketplace on the basis of statistical data (Schoeffler, 1980a). But statistical validity is not the only factor that distinguishes natural laws from those of the market. The major distinguishing characteristic of any natural law seems to be the specification of the conditions under which it is true. Empirically established correlations between different factors (in other words, are derived from the phenomenological observation of the world, the inductive method), are usually supported with some deductive theory. Meanwhile, such conditions are not stated explicitly for the laws of the market uncovered by the PIMS project, although they do represent causally determined relations rather than mere correlations.

Strictly speaking, no physical law specifies all the conditions under which it is true, and physicists often have to resort to some kind of an idealization of real life conditions (this ranges from the laws governing the behavior of ideal gases to those of gravity). But, nevertheless, conditions under which a given law is valid are always specified.[3] Engineers see these laws in a different light. For instance, in applying the laws of ideal gases to a given gas under given conditions, they are forced to introduce adjustment coefficients into the equations. Here, they may use either intuition or some set of heuristics worked out previously for guidelines. And no matter how much these coefficients deviate from the initial base, the latter is always specified in precise terms.[4]

Laws of gravity reflect reality to an extent that prohibits engineers from making any modifications. The laws are valid both on earth and on the moon. No matter what engineers are making, a car or an airplane, they must accept these laws unequivocally.

The authors of PIMS realize that so-called market laws merely represent statistical averages and must therefore be modified when applied to each particular case. But, as is usually the case, knowing these general laws helps one in choosing the proper course of development. For instance, Schoeffler (1980b) believes that market laws discovered by the PIMS group explain about 80 percent of all cases of profitability and unprofitability.

But the claim that these market laws are analogous to the laws of natural sciences is not totally correct until all the conditions under which they are valid are specified. The result, which accounts for about 80 percent of the firm's performance, is only true on average. In each given the deviations from the norm may be quite large. But what is important here is that the manager using these laws is not familiar with all the conditions under which these laws were formed, the starting point for his analysis.

Finally, it should also be noted the PIMS statistical data applies to a rather rapidly changing system and may be insufficient to make any final conclusions, if only because it covers a rather short period of time. Furthermore, businesses participating in this study may not be representative of the market in general. (It is unclear whether these businesses were picked selectively or randomly from all those wishing to participate in the Harvard project.)

The theoretical part of the problem is still to be completed although the PIMS authors have already taken the first steps in this direction. Indeed, their work is not limited to statistical verification of the quantitative contribution of each individual factor; it also includes an analysis of these factors.

For instance, in Ravenschaft and Scherer's (1982) work, the effect of a material parameter (R&D) on profitability was analyzed in conjunction with a relational parameter (market share). The greater return on investment in marketing as compared to R&D was attributed to the relative immediacy of the results — R&D starts paying off only after four to six years. At the same time, analysis of R&D expenditures has indicated the need to reexamine current accounting methods. It was discovered that investment in new product development pays off much faster than investment in the development of new technologies.

In one of his works (I was unfortunately unable to find any reference to it), Schoeffler presents an interesting analysis of the investment intensity parameter and reasons for its inverse relationship to profitability. The author notes that the cause of this lies not in such a seemingly obvious thing as a high ROI denominator (in case of high investment intensity); taking the amount of investment in proportion to the ratio of sales to profits does not change our general conclusions on the qualitative effect of investment intensity on profitability. It was also shown that investment intensity has no effect on another important parameter, market share; that is, they are not strongly correlated. Therefore, the authors concluded that in order for investment intensive industries to operate efficiently, they have to have high capacity utilization, which is not easy to achieve.

Now I want to comment on the practical applications of the PIMS model focusing on the aspect of the time dimension in the utilization of the model. The market laws discovered by the PIMS model apply to a time period of three to five years. I am not sure why the PIMS authors chose to limit their model's horizon to this particular period. Perhaps they were led by purely pragmatic considerations, knowing the actual horizon set by corporate planners. As representatives of the executive branch, they are primarily responsible for handling current problems, leaving them limited time to consider long-term perspectives.

Whatever the reasons for the particular horizon set by the authors of PIMS on the validity of the laws, I want to discuss these reasons in greater detail. I hope my discussion will help clarify the conditions making the authors' claim of three to five years a reasonable estimate, develop procedures to turn this model into a practical tool for the businessmen, and find ways to make the model applicable to other time intervals. There are at least two ways to approach this subject. At one extreme is the problem of the near future, say three to five years, and at the other is a far more distant future, say ten to fifteen years.

In the first case, the primary use of the PIMS model is to evaluate firm performance based on the weight function where the material parameters are measured in terms of current prices, and relational parameters are evaluated subjectively by the firm's leaders. In the case at hand, the aim of the PIMS model is not to generate objective decisions on the proper course of development, but to enrich the intuition of the decision makers in assessing relational parameters as well as the above mentioned unmarketable assets. Here, the significance of the model lies in its being the first step of an iterative process undertaken by planners in their attempt to determine the profitability of different business ventures. The model may, for instance, focus the planners' attention on the most significant factors affecting profitability, although their own intuitive priorities may be quite different. Discrepancies between the weights assigned to various parameters by the model and their own estimates may stimulate them to substantiate their intuitive notions by analyzing the reasons for those discrepancies.

Nevertheless, I think the authors of PIMS overstress the objectivity of the model's findings and truly underestimate the role of the subjective factor. Although objective factors may, on average, actually account for 80 percent of the variation in profit, subjective factors such as a manager's personal abilities may cause large fluctuations. Here, it is not only a manager's personal drive and leadership qualities that must be taken into account, but also the subjective evaluation of different factors that in turn depends on personal understanding.

These considerations affect the potential of a corporation and must be taken into account both by the existing leadership and when the business changes hands. The personal role of the executives in the continuing development of a corporation may vary greatly. Their presence can retard development, in which case they are removed, or it can be a significant asset, in which case they are encouraged to stay even when the ownership of the business changes.

Another, and possibly much greater, benefit can be derived by using the PIMS model in long-range planning, say 10 to 15 years. Here, the major problem is in formulating the final goal — the state to be reached at the end of the planning period, that is, specifying the structure of the desired outcome (using the terminology from the theory of control, we have to state the conditions of transversality). We must avoid the so-called Pyrrhic victory. To be useful in long-range planning, the model would have to incorporate a number of additional features. The position of a firm at the end of a 10-to-15-year period is based on the structure of its input-output (the material parameters) and their relationships.

As long as market prices can be used to estimate the value of material parameters, the weights of different indicators affecting firm performance can be evaluated using these market prices. This is precisely the approach used by the authors of PIMS. But the situation is very different when we are dealing with a more distant future. The information about material parameter prices is unavailable, and it may even be safe to assume that the mechanism for estimating them is of an indeterministic nature. Therefore, we are forced to introduce a new kind of price. The problem of computing unconditional prices does not enter the PIMS project, for its authors have chosen to operate their model on the basis of current prices.

In summary, the greatest contribution of the PIMS model is in its pioneering attempt to quantify the many relational parameters affecting firm performance. Operationally, this model is very similar to my own concept of potentiating, both in its emphasis on the firm's position and on relational parameters. However, the PIMS model focused on the econometric analysis of the problem under a rather narrow theoretical framework. Consequently, the model missed the need to introduce unconditional prices and neglected the subjective factor, particularly in finding the values of relational parameters.

NOTES

1. The set of individual reflexes can be viewed as heuristics because each one behaves independently. See Katsenelinboigen, 1989.

2. Here the criteria used to judge the work are the same as those used by mathematicians in verifying the proof of a theorem. Prof. Boris Moishezon has pointed out to me that mediocre mathematicians often try to attack a new mathematical idea by finding a mistake and declaring an entire theory invalid. Mathematicians of a higher caliber are more concerned with helping the author overcome the difficulties while preserving the novel ideas present in the work.

3. As we know, even in mathematics, the verification of all the conditions making a given theorem true represents a difficult problem. If I understood Prof. Moishezon correctly, many theorems "proved" by the Italian school of algebraic geometry at the end of the nineteenth century were either found to contain mistakes or to be incorrect altogether because they failed to provide a rigid set of supporting axioms. Still, compared to other sciences, mathematics has been more successful at specifying all the assumptions leading to a given conclusion.

4. With characteristic wit and paradoxical way of thinking, Russell Ackoff one time noted that engineers are less precise in their calculations than economists. An economist will never double or triple his initial estimate of the staff requirements, but it is not uncommon for an engineer to double or triple the initial estimate of the safety margin of a bridge. However, it seems computations performed by engineers are more exact. The two professions have absolutely different starting points. Engineers choose to modify the initial results because their calculations are based on ideal conditions. Economists perform their calculations based on experience, which is then adjusted. I believe the final specifications of bridge strength set by the first builders-craftsmen were not much different from the results of their draft calculations.

16

Firm III: Positional and Combinational Styles of Management

The last chapter dealt with several general ways of fully integrating material and relational business performance parameters. Now I want to examine what varieties of the positional style, and especially its combinational version, have been used by two eminent U.S. scholars and consultants, Russell Ackoff and Peter Drucker. Ackoff has come to his theoretical conclusions largely through his own practical experience in the field; Drucker, in his conceptualizations, has relied mainly on other people's observations.

I had to sacrifice somewhat the logical flow of my exposition to present a more rounded picture of these scholars' views by mentioning briefly an entire array of styles that they advocated, not only the principal varieties of the positional style.

CHOICE OF MANAGEMENT STYLE IN THE WORKS OF RUSSELL ACKOFF

Russell Ackoff has made use of different styles to solve business problems, and whatever style he chose to work with, the solutions were always innovative. Let me illustrate by citing Ackoff's approach to the reactive and connective-combinational styles. Conceptually, the technique of a reactive style is fairly simple and uses statistical data to establish correlations. Ackoff's results, however, have been quite unconventional and even counterintuitive. He showed, for example, that under certain conditions reducing expenditures for advertising could lead to increased sales.

One well-known business practice is the use of a heuristic rule to determine expenditures on advertising as a percentage of total revenues. In the 1960s, a project by Ackoff and a group of colleagues (1975) involving a large beer producer, Anheuser-Busch, put this rule to a test. The task of this group was not simply to establish statistical correlation between advertising expenditures and consumption, but to explain the reasons underlying the impact of advertising on the consumer. "We believed that what these analyses really showed was that most companies forecast next year's sales quite accurately and then set their advertising budgets as a fixed percentage of predicted sales. In other words, forecasts of increased sales produce increased advertising" (p. 3).

The study of the impact of advertising on beer consumption was predicated upon the following hypothesis: the relation is described by an S-shaped curve having a saturation point, beyond which additional advertising can actually be counterproductive. This hypothesis was subjected to statistical verification where the structure of advertising itself was examined. Naturally, convincing management to assume the risk of reducing advertising expenses under stiff competition in order to increase sales required considerable imagination on the part of the research group.

This study was quite beneficial, or as the authors of the quoted article wittingly note, at least it did not make things worse (perhaps, paraphrasing a similar remark by a traveling salesman talking about an amulet that had a cross, star of David, and such, in Stanley Kramer's movie *The Ship of Fools*). As the project unfolded, the sales of Anheuser-Busch's major brand of beer, Budweiser, doubled, and advertising expenses per barrel of beer were reduced by 58 percent (p. 12).

In time, Ackoff outgrew the connective style characteristic of operations research, but his attachment to the combinational framework has remained largely intact. In setting goals he took account of relational parameters, but he did not evaluate them per se, preferring to see them only as hints or guideposts to indicate the direction in which increased profits should be pursued.

In one relatively simple case of Ackoff's (1970) combinations, the "guideposts" were the relational structural parameters expressing varying profitability of a given company's products. The specific goal was set in terms of material parameters: to eliminate as many unprofitable products as possible. The resulting operational plan was quite sophisticated: to concentrate the incentive system for sales on profit rather than volume.

In this case Ackoff drew attention to one structural parameter — relative profitability of different product lines. The associated relational parameter, variety, was left out of account. The chosen structural parameter, moreover, did not receive an independent evaluation but was used instead as just a guidepost to indicate the possible directions of profit maximization efforts.

Ackoff's approach is obviously not without merit. A direct link-up of current activities with the final goal of profit making is often desirable. Still, a word of caution is in order, and managers should be warned that this way of pursuing profits, however attractive in the short run, could result in the worsening of certain relational parameters and ultimately hurt the company. In the present case, the growth of the material parameter of profit was achieved at the cost of narrowing the breadth of product lines, that is, at the cost of such an important relational parameter as variety. Potentially, the resulting hole could be exploited by competitors. If, nonetheless, the decision is made to proceed with a variety-reduction plan, steps should be taken in advance to prevent future competitors from encroaching on one's territory through these holes.

A number of U.S. corporations have reportedly had financial trouble because they did not produce all the equipment types demanded by their customers. The holes were the result of deliberate policy, not oversight: it was simply inefficient to produce certain kinds of machinery in small amounts. Large foreign producers of all relevant types of equipment entered the U.S. market through the holes by initially offering U.S. dealers good discounts on the equipment they did not have. In due course, when the dealers got used to their high profits and learned the corporate ways of foreign companies, the latter began offering them good discounts on all types of equipment. The stateside companies, as a result, soon found themselves on the brink of bankruptcy.

Over the years, Ackoff has become more appreciative of the positional style, especially of its strategic version. A set of relational parameters used by Ackoff can be conveniently divided into two categories: the external ones that describe the position of a company in the market and the internal ones that characterize the intra-corporate activities.

Let us take an example pertaining to the first group of parameters — a case as told by Russell Ackoff who headed the group consulting Company C.[1] Company C had about 18 months to bankruptcy. The bank said if the company did not change the positive cash flow it would put them in Chapter 11, liquidate them, and take their money out because Company C in this time did not service this debt in five banks. The first thing that the leaders of Company C did before we came over was to ask two questions. The first one: "How long would it take us to redesign our product line to be competitive with Company A and to be able to produce at a price comparable to them?" At that time Company A was a giant and was selling better products than Company C was making and at a price that was lower than Company C's costs. Company C, in order to maintain market share, was selling at a loss, and that is why it had negative cash flow. So Company C learned it would take about eight years to redesign the product line and the production process and implement it. Then Company C said: "Well, we will change only the production process, but we'll try to produce it less expensively so that we don't lose money on every sale." That is why the second question was: "How long will it take us to change the production process?" And the answer was: "Three years!"

The only alternative available to Company C was to prepare the company for sale and try to sell it. To sell it to somebody else could provide the capital to keep it alive and invest in improvement. That is what the staff of Company C recommended when the new leader, Mr. W took over.

We took Company C's leading people to a university for a week and said to them: "We want you to assume that Company C was destroyed last night. Company A is still there. Company B, another giant, is still there, all the other companies are still there. Design a company that you would replace it with, anyone you wanted." They objected. There was a terrible fight that day. They said, "It's a waste of time," and "What's the point; we can't do anything. Why should we talk about the ability to do anything?" Mr. W was very tough. He said: "If you want to draw your salary, you're going to do this. You don't have to do it, but you don't have to draw your salary." Basically it was his position. So they decided to do it.

By the second day they started to get excited because they discovered something about themselves. What they discovered was they had tremendous knowledge, which they had never been able to use because, in the context of their company, they could not use what they knew about their competitors. They began to redesign and began to use that knowledge. At the end of the week, they had two major realizations. The first was they knew how to design a company, which if it existed would compete very, very effectively with Company A and Company B.

The second thing was they did not know how to build it. Mr. W said: "I didn't ask you, 'Can we get there from here?' That's not the way we're going to do it. What I want you to do is work backward from there to here, not from here to there. There are about 25 other companies out there in the world that are in the same boat we're in. Suppose you could put them together, any number, any one. What's the closest approximation you could get from the ideal?" They thought that was great. But who the hell wants to form a joint venture with a company that's almost going to go bankrupt? It doesn't make any sense. Why? I replied: "I don't know why but after we're done we'll know why." So they took a second week and discovered four companies; Company D from the Country F, Company E from Country K, Company L from Country Y, and themselves. And again they were amazed because it was so close to the ideal.

The basic question for this merger of four companies was how to compete with Companies A and B. Companies A and B were the only two companies in the world that produced a complete product line, so it was possible for a dealer to have all its products from the same manufacturer. This holds a distinct advantage in terms of spare parts and in terms of service. Each of the other companies dealt with only part of the product line. That is why we began to design a company to produce everything. Company C actually made a matrix in which on one side one can see every piece of related products available anywhere in the world, every single piece. Then they put on the other side what they were producing so that they could see where they were. They found that Company D made equipment they did not make. Company L made equipment that either one of the other two made, and Company E made equipment that none of the other three made. They overlapped on some of the equipment. For example, there were certain types of products that both Company C and Company D made, but together they did not make all the sizes available on the market. They redesigned their joint product line so that they had representation in each size.

But to have a new design is not enough.

"Look," Company C's managers said, "there is no way we can get those companies." And I said, "I know, but we're not sure. Why don't we go talk to them? What have we got to lose?" Mr. W said: "Well, show them the design because that's one thing they've never seen and see how they react to the design." So Company C's managers agreed. Mr. W loaded the reluctant managers into the company airplane and went with them to see Company D, which was part of Company D-D. The leader of Company D-D said to them: "We are so excited by the idea. You know, we've had six offers for Company D in the last year. We turned them all down. Do you know why we turned them down? Company D is one of the first companies we ever bought. We would be ashamed to turn it back in a bankrupt form. But all the people who want to buy it want to liquidate the company, take its assets, convert it into cash, and walk off with the cash. Your design is the first thing we've seen that offers hope for the survival of the company as a company. We will sell!" But we said: "We can't buy it; we don't have any money." They said: "You don't need any. We'll do it with exchange in stock." Company D-D got 10 percent of Company C, and Company C got Company D.

Then they went to Company E. It happened we were working at Company E at the same time, so we could set up that meeting for them. They met with the CEO and the president of Company E. He was excited but for a different reason. He said: "You know, we have never been able to successfully sell our product but you are very strong in this area. On the other hand, you are not very strong in some regions, and we're strong in others." And he went around the world. He said: "If you put these two companies, Company C and Company E, together we've got the strongest marketing organization of any company in the world in our business. But there is no way we can get Company L's management, or the management of the other companies, and we can't even get the CEOs of these companies to work together harmoniously. The cultures are too different, the value systems are different, the criteria are different, they cannot work together."

Mr. W said: "You are probably right, so why don't we test it. Let's pick five people from each of the companies, put them together into a team, and have them design the closest approximation using the three companies for the design I just showed you." Company E's president agreed.

In April 1985 that team was created. There were five managers from Companies C and D combined, five from Company E, and two of us from the university. So we had a team of 12, and we went to work designing. By September we had designed that business to the finest detail. We had five factories, knew which wholesalers would be retained, what the logo of the corporation would be, and so on. We took it to the board. It was unanimous, and the cultural problem never arose in the group. They unanimously accepted the proposal. One year later, all the legal work had been done, and they created the C, D, E Corporation. It became the third largest company in the world and has been profitable since then.

Company L had just completed a five-year contract with another Company G in which they would market one of that Company G's products. Company G would also market its own product. They could not break that contract. So they said: "Hold it until our contract is done, and then we'd like to discuss it with you." In the meantime Company C looked at alternatives in case it did not work out with Company L. Company C had two fallback positions because it wanted a product line as complete as those at Company A and Company B. With the merger of Company C and Company D there was only one particular area where the united company did not have products for a complete line. Company L was to supply that. But there was a company that produced the same product. It was not as good as Company L's, but it was better than not having the product.

As Ackoff's story demonstrates, Company C could not have avoided bankruptcy by traditional means of cutting costs and increasing revenue through an improved product or a better production process simply because there was not enough time. The solution, based on a relational parameter of monopolization, consisted of changing the market position of the company.

Indeed, the key problem faced by Company C had to do with its being an underdog that could not compete successfully with oligopolist Companies A and B. The challenge, therefore, was to turn Company C into an oligopolist, too. A step-by-step strategy was developed, with goals for each stage expressed in terms of relational parameters. The final stage was to begin with the realization of an intermediate goal of convergence (a relational parameter) because a new oligopoly could be created in short order only by merging several

underdogs. Such a nontrivial solution in a situation such as Company C's impending bankruptcy was possible only if all members of a would-be oligopoly could materially gain from it. The gain, in turn, could be secured through filling in the holes in the product span left by individual companies, that is, through their manufacturing jointly an entire array of relevant equipment, another relational parameter. The new design of Company C was geared toward that objective.

Ackoff's concept of "ideal-seeking design" was the crucial methodology behind Company C's reorganization. Paradoxicalness, which is Ackoff's forte, together with his extraordinary ability to combine philosophical and practical modes of thinking, found a happy match in that instance in the similar gifts of the company's new president. Such collaboration again proved to be extremely effective.[2]

Ackoff applied his positional thinking to the problems of several other companies. The strategic emphasis there was on relational parameters known as "human factors." Ackoff and Deane (1984) show that in an effort to boost efficiency managers often choose to reduce the value of the material parameter known as employees' wages and salaries. The move makes sense if it does not simultaneously entail lowering the value of a structural parameter expressed as a ratio of unitary output to unitary compensation. It might prove possible, in situations of increased unemployment, to improve the value of this parameter by cutting compensation, but the effect is likely to be only short-lived. As soon as the employment situation gets better, the employees will stage a comeback with a vengeance! If, however, the management finds a way to improve the parameter in question without cutting compensation even when unemployment is high, the effect is bound to be lasting and will not dissipate when the employees feel more secure about retaining their jobs.

The idea, then, is to change radically the management-employee relationship from adversarial to cooperative by explicitly setting the overall strategic goal of achieving cooperativeness — a relational parameter that has an independent value irrespective of prevailing conditions.

These ideas were applied to the situation at ALCOA Tennessee Operations by an Ackoff-led group at the University of Pennsylvania, Bush Center. The company was losing money and even faced the threat of closing. It was clear that the value of such material parameters as unitary labor costs and product quality had to be improved. But it was also clear that such relational parameters as operational conservatism and deep-seated management-worker hostility stood in the way of attaining that goal. From the outset the Bush Center consultants showed a great deal of ingenuity just to get the management and the union people to sit down together at the negotiation table. The success of the project exceeded all expectations.

Another successful case concerned a reorganization of a supermarket chain based on worker participation in management decisions and a changed nature of employer-employee relationships (Ackoff & Deane, 1984; Mayer, 1986).

An extraordinary synthesis of theory and practice achieved by Russell Ackoff has had a profound impact upon his students. Let us look at the case of reorganization of strategic planning at Armco Steel, Latin American Division, carried out jointly by Julio R. Bartol and Ali Geranmayeh, a former student of Ackoff (1989).

The starting point was Ackoff's concept of ideal-seeking design. In thinking about design implementation, the authors decided against following the prescriptions of the theory of "disjointed incrementalism" or those of such rational and comprehensive schemes as PERT/CPM. Instead, they chose my idea of potentiating and fleshed it out through an original five-stage spiral-like plan based on a strategic style and the formulation of different goals appropriate to each stage. Other ways of ordering the stages besides the one used by Bartol and Geranmayeh are also possible.

Stage 1 called for gaining some well-defined business objective (the objective, appearance to the contrary notwithstanding, was not a combinational but a positional one, as the subsequent exposition of the spiral-like, multistage development plan will make clear). At Stage 2 the positional goal was to expand distribution, that is, to achieve market divergence with unchanged technologies and product lines. Stage 3 involved a positional goal of expanding the product catalogue, that is, product divergence. At Stage 4 the positional goal had to do with finding complementary materials and developing new product components, specifically replacing a number of steel parts with plastic ones. Stage 5 was the same as Stage 1 but at a higher segment of the spiral.

Stage 5 brings out in bold relief the positional goal established at Stage 1. The final task was to find a complete set of functionally identical goods and services in an industry (for example, grain storage bins, water tanks, and irrigation pipes) regardless of whether or not they are being produced by the given company. The approach is revolutionary because enterprises, locked as they are into existing technologies, tend to produce specialized products and do not usually think about a complete set of goods and services needed to perform a certain function. One of the secrets of success at Armco's Latin American Division was precisely in making explicit the importance of completeness as a positional parameter.

The company, for example, used to manufacture grinding balls for horizontal rotating mills that grind large mineral-containing rocks. With a goal of a complete set of functionally integrated goods and services in mind, Armco began supplying its client mining firms with other types of machinery and equipment, including computer software for sophisticated technological processes. Eventually, it was able to develop the most efficient rock-grinding methods in the world.

COMBINATIONAL-AGGREGATE STYLE

The definition the combinational style stipulates that the material parameters set as goals must be evaluated in terms of unconditional prices. However, we do not see this type of price being used in practice. Rather, the material parameters are evaluated based on current prices, so the classic version of the combinational style manifests itself in practice as a combinational-connective style. Nevertheless, the methodology underlying our analysis of combinational style is invariant with respect to our choice of the base prices assigned to target material parameters, so we shall proceed under the general banner of combinational style.

Before we address the classical version of combinational style, which encompasses a very limited number of material parameters (often a single parameter expressed in physical terms), let us look at one subclass of this style, combinational-aggregate style. As the name suggests, the goal is delineated by a cluster of material parameters represented in an aggregate form. The aggregate, in turn, may be given either in physical terms, for instance, the weight of machinery produced (this indicator refers to the total weight of heterogeneous heavy machinery produced singly or in very limited amounts), or in monetary terms.

One feature intrinsic to this style was to change profit from the leading unified differential indicator to one subservient to one of its two constituent components — expenditures and sales volume. I have coined a term for this approach to firm performance — the principle of asymmetry — in contrast to the classical method of evaluation (the symmetry principle), which is based on the parity of all the constituent components of profit; that is, they are not assigned priorities.

The asymmetry principle with respect to profit means just that: some components of profit are given higher priorities than others relegated to subsidiary roles. The principle of asymmetry manifests itself vividly relative to the two major facets of profit — revenues and expenditures. It implies that the firm aims at the growth of sales with all other parameters contributory to this aim or that the firm keys in on reducing the costs and demands other parameters to serve this cause.

It is well established (although Peter Drucker [1954] casts some doubt on this) that William Baumol (1967) pioneered the criticism of profit as the sole criterion for evaluating firm performance and authored the concept of formalized correlation of profit with supplemental indicators. Baumol raised the issue that alongside profit there ought to be an independent performance indicator like sales growth. "I believe that to him [a businessman] sales have become an end in and of themselves" (p. 46).

In his research Baumol detailed why profit is insufficient, why alongside it there has to be an important indicator like sales growth.

> There are many reasons why the businessman should show such concern about the magnitude of his sales. Declining sales can bring with them all sorts of disadvantages: There is reason to fear that consumers will shun a product if they feel it is falling in popularity, thought their information on these matters is certainly often spotty. Banks and the money market will tend to be less receptive to the desires of a firm whose absolute or relative sales volume is declining. Perhaps even more important in this connection is the very real danger that firms whose sales are falling will lose distributors — a major marketing setback. Management is also not unmoved by the fact that in a declining firm personnel relations are made much more difficult when firing rather than hiring is the order of the day. The firm that declines (or remains small when others expand) can lose monopoly power and the power to adopt an effective competitive counter strategy when one is called for, and may become more vulnerable to a general deterioration in business conditions. For all these reasons the executive may reasonably conclude that maintenance of as large a sales volume as possible is the only way to succeed in business (pp. 45–46).

With this in mind, the optimality problem characterizing firm performance might be represented as an effort to maximize sales under the constraint of a minimum profit. Of course, between the drive to increase sales and the size of profit there could be conflicts. Establishing a compromise between them is complicated.[3] Baumol, however, proposes that the difficulty in setting up a measure for sacrificing profit for the sake of sales and the need to somehow externally establish this measure as a constraint is characteristic only of short time intervals. As for long periods, profit could become not an external constraint, but an unknown variable.

Thus, according to Baumol, profit in all cases plays a subordinate role. At each point in time, it will be possible to find an effective profit measure from the point of view of a criterion for the firm expressed dynamically not simply by the drive to maximize sales, but by the impetus to maximize the rate of sales increase.

From the methodological standpoint the attraction of Baumol's concept is evident. When there are a large number of independent variables characterizing firm performance and it is quite difficult to integrate them, especially via such a catchall indicator as profit, it is tempting to solve the problem by establishing a dominant indicator and through it determine a value for all the others. However, this methodological solution is quite insidious because, there are many parameters, a notable portion of them (if not all) are independent variables. The methodological error committed by such powerful thinkers as Marx and Freud was that they tried to find a dominant parameter in their fields through which to express the value of all other parameters. For Marx this parameter was the production of goods; for Freud it was sex.

One must be acutely aware of the deceptiveness of solving the problem of measuring the state of multiparameter objects by using dominant parameters. In this connection Baumol ought to be given his due. He does not claim his performance criterion should be the only one. As he himself notes, it was important to show that the profit criterion is insufficient (p. vii).

If we look at the works of Michael Porter (1985) from the standpoint of the asymmetry principle, we see that, in contrast to Baumol, he focuses on the other component of profit, cost reduction. Unlike Baumol who presents his material component (sales growth) unalloyed, Porter supplements his material component of cost reduction with the relational parameter of differentiation. "The fundamental basis of above-average performance in the long run is sustainable competitive advantage. Though a firm can have a myriad of strengths and weaknesses vis-a-vis its competitors, there are two basic types of competitive advantage a firm can possess: low cost or differentiation" (p. 11).

The main reason to adopt cost reduction and differentiation is purported to be the following: "After several decades of vigorous expansion and prosperity, however, many firms lost sight of competitive advantage in their scramble for growth and pursuit of diversification" (p. xv). The argument for stressing the expenditures component of profit is just as strong as for its other component — the growth of sales. There is no need to dwell on this topic for the arguments are rather trivial.

Reasons for the firm to focus upon one of the two components are many and diverse. It seems one of the main factors is the ratio between the amount of time expended by the firm's leadership (who must focus their attention on some specific aspects of profit) and the derived effect of the proposed initiatives. Consequently, in real life the leadership picks some single component of profit and proclaims it as the cardinal aim, focusing all effort upon its realization. At one time or another, one component of profit may be more fashionable than the rest, which, of course, does not mean that all firms must blindly follow the dominant trend (just as stylish women choose not to stick blindly to the ruling fashion, but they always keep the new in mind).

An interesting example in this regard is Lee Iacocca (1984). After becoming president of the Ford Company he placed priority for the upcoming several years on reducing expenses. Before assuming that post he had gained a high profile in the company by directing attention toward increasing production by building new cars that were attractive to customers.

CLASSICAL COMBINATIONAL STYLE

I would venture that the one classic work devoted to combinational style is the book by Peter Drucker (1985), one of the founders of modern management theory. The book explores the interrelation between management and innovation. Certain parts of the book appear to be a product of common sense. Other parts expound guidance typical of popular textbooks on management. There are plenty of anecdotal stories about successful and unsuccessful businessmen. But all this is a by-product of a more serious endeavor. The book makes important strides in conceptualizing the way business leaders formulate problems, or more precisely, the way they search for goals. Innovation, according to Drucker, must be goal-oriented or purposeful (p. 290).

Drucker's book also focuses on the methods employed by business leaders to attain the set goals. The author aims to show that consistent development of the economy can be achieved only by combining management skills and innovation. Here, to keep our discussion from going astray, I am not going to dwell on the methods of goal attainment explored in the book.

As I have mentioned above, setting goals narrowly defined as specific material parameters is the hallmark of the combinational style. Such goals can be set in a twofold manner. The key task under the first approach is to uncover the consumer demand function and then set the need-fulfilling goal, that is, a product that fulfills this function in an optimal manner. This represents goal setting initiated at the end; that is, it hinges on the function. The second way is to start at the beginning; in other words, proceed based on the structure of the existing potential. One example of this approach toward creative ventures is discussed in the work of James Adams (1980).

Drucker's work illustrates both types of approaches to goal selection. According to Drucker, the initial stage in the process of goal selection is understanding "the users — their utilities, their values, their realities" (p. 252). He is less concerned with such environmental factors as suppliers and competitors. For instance, his emphasis is not on shielding oneself from competition, producers of similar products, but rather on gaining the upper hand by launching a successful assault on the consumer. To do it, Drucker suggests that the producer has to identify the function fulfilled by a given product (or service), rather than the product itself.

Drucker illustrates this particular concept using the case of Gillette shavers. In the 1870s, following aristocratic ways, the middle class began to shave, and it faced an unpleasant choice between using a barber razor, which was quite dangerous, or going to a barbershop, which was relatively expensive. A safety razor that cost five dollars was hardly marketable, for a visit to the barber was only ten cents. Gillette sold its razor at a small fraction of what it actually cost them to manufacture it. But it priced its "razor-specific" blades, which cost the company less than a penny, at five cents. This proved a successful strategy because the customer priced a shave, not the product sold.

Before further discussing the methods of goal selection proposed by Drucker, I would like to note in passing one essential feature of combinational, as opposed to positional, style. The latter explicitly promotes creating a good position (incorporating all or a number of relational parameters as well as material parameters). The combinational style scrutinizes the existing position, attempting to uncover relevant relational parameters that can help identify some specific material component, which can then be set as an intermediate goal in the firm's overall path of development.

It is not by accident that the last problem, which was designated above as the second approach to goal selection, commands the greater part of Drucker's book (a total of 277 pages). Let us pursue this problem focusing on the various relational parameters introduced throughout the book.

It seems we can match Drucker's seven sources for uncovering relevant parameters with my own classification of parameters characterizing the state of a system: industry and market structures and demographics correspond to primary structural parameters of a position; incongruities, process need, the unexpected, and new knowledge equal the primary characteristic parameters of a position; and change in perception matches conjugate parameters of a position.

These sources are all mutually intertwined, and yet each one is peculiar. The author illustrates each source with an array of examples borrowed from real life operation of market economies, especially the U.S. experience. Drucker treats the environment shaping these sources as a given. The leader of an organization need not shape the environment, nor even know why the environment has changed, why the situation is what it is (p. 50). In other words, Drucker contends that the leader need not bother to learn how the position of the environment is shaped. The crucial task is to derive the greatest benefits from the ever-changing position.

At this point, I will offer a more in-depth analysis of the aforementioned sources, starting with one important subclass of primary relational parameters that define a position, the structural parameters. Structural parameters pertain to social as well as demographic entities. Sometimes uncovering these parameters is simple (just by looking at the statistical data describing the respective phenomenon), but, at other times, it entails highly nontrivial methods. In other words, the problem is twofold: we may search for new structural parameters other than the widely recognized ones, or we can take known parameters and trace their history.

One example illustrating the first case is the famous fiasco of Ford Edsel, manufactured in 1957. The flop was caused by a wrong social orientation of the new wonderfully designed and manufactured automobile. Edsel was intended to fill the gap in cars geared toward special social stratum (lower, middle, upper, and their internal isomorphic subdivisions). However, this traditional stratification was insufficient. Successful marketing called for another complementary classification of the population based on the concept of lifestyle segmentation. Having identified, based on the new stratification, the missing link between automobiles being offered on the market, Ford designed and produced the commercially triumphant Thunderbird.

Drucker's book, especially Chapter 7 on demographics, presents numerous examples of successful firms. These cases embody the evolution of known structural parameters toward the relevant ones that could help identify the firm's combinational objective. One of the then minor shoe chains capitalized on the baby boom by opening stores catering specifically to teenagers. It then expanded into clothing, and, by the time others invaded the market for this narrow age group, it shifted its sales to accommodate the slightly older group into which the baby boomers had by then grown.

Another subclass of relational parameters, characteristic parameters, is also used to guide the firm's search for a goal. Let us consider a case where there is only one such characteristic parameter, for instance complementarity. As TV began to encroach on the U.S. public, a number of publishing houses in the United States spotted this particular parameter and managed to convert it to great success. With the advent of TV in the 1950s, everyone thought that people in the United States, which had never been a book-buying country, would almost stop buying books altogether. In reality, book sales have far surpassed the relative growth of every other indicator, be it income, the size of the book-reading population, or people with higher education.

Many examples of how firms set goals describe a goal that incorporates both structural and characteristic relational parameters reflecting firm performance. The case of British padlock export to India described by Drucker is a case in point. One hardware company's major product line was cheap padlocks exported to British India. Rising incomes in India in the 1920s brought a considerable decline in the sales of this item. The company improved the lock, improving its quality with minimal increase in price. Contrary to company expectations sales plummeted even more. So, what happened? In India in rural areas a padlock's use was purely symbolic. No thief would dare tamper with a padlock, so peasants never bothered saving the key. They were able to open the old cheap padlocks quite easily with a pin. The better lock did not yield so easily. At the same time, the growing middle class in India's cities wanted a lock sturdier than the improved version sold by the company. The firm began to decline and eventually went into liquidation.

The competitor recognized the situation and reaped considerable profit by marketing two types of padlocks: one that opened easily with no key (actually cheaper than the old padlock, but with a much higher profit margin) and another sturdy lock with three keys (twice the price of the old lock but still with a higher profit margin). This example reveals that the goal of the firm was to adjust to a shift in the structural parameter of stratum ratio — growth of the middle class — as well as recognize such characteristic parameters as closeness — the use of the lock in rural India as a cultural symbol of a house being locked and as real means of deterring thieves.

Obviously, a person of Drucker's caliber cannot confine his analysis of the methods of goal selection to a combinational style. The author resorts to other styles of decision making, particularly the positional style. Drucker, himself, does not distinguish among the various styles; in fact, the use of such terminology is foreign to his approach. But he does call our attention to the fact that a leader has set a certain goal as the cornerstone of development, rather than just react to a situation at hand.

Although Drucker's book does not abound with examples of other styles of decision making, one example uncovers the problems in trying to implement the positional style. The executives at Macy's department store, which specialized in clothing, were concerned with the rocketing sale of appliances at its store. Although the sale of appliances contributed to the volume of sales at other departments, the share of overall sales at the fashions departments dropped considerably. To preserve the image of the store, they tried to increase the sales of fashion, "to restore the normal ratio," and, when this did not work, the management took steps to decrease the sales of appliances. The department store continued to flounder until new management modified the policy and reversed the downward drift. Another large department store in the area restructured its operations to derive the most benefit from this unexpected boom in appliance sales.

The story makes it clear that the dilemma facing the executives at Macy's was the choice of a characteristic parameter, concentration versus dispersion. In other words, the choice was either to focus the leaders' efforts on clothing or diffuse their time on a wide array of different products. The leaders of this fashions establishment faced another tough choice between sticking to their traditional specialization or embracing innovation by introducing a new line of products (another characteristic parameter, tradition versus innovation).

Putting all the relational parameters, which Macy's leaders must have taken into account, into the spotlight might lead us to conclude that Drucker's verdict, that the leaders picked a wrong strategy by closing a highly profitable appliance department, is somewhat of an oversimplification. It seems that this strategy represents a typical positional sacrifice — the decision to sacrifice profits for the time being and assign high value to certain characteristic relational parameters.

The first relational parameter to consider is traditionality. The numerous examples in Drucker's book indicate that entrepreneurial leaders are highly mobile; they are willing to break from the entrenched mold. They are not strongly tied to a tradition; i.e., they are less Markovian.

However, the problem of replacing the traditional orientation of a firm is not a simple matter. The leader could approach the problem by shifting the burden of being a pioneer upon other firms that are willing to take a chance by being innovative and breaking away from the entrenched tradition. Then our more conservative leader (as well as other players in the same field) would follow the pioneer's footsteps imitating the tested business practices.

Perhaps, it is too much to ask that every leader of an organization be a pioneer. A firm that has specialized in a certain product or service, and has done so successfully, might be well advised to retain it traditional profile. Perhaps, it would be detrimental for such firms to scatter the leaders' attention and effort over an expanded sphere of activity. Perhaps, the situation being what it was, it made sense for Bloomingdales to assume the pioneering role and expand the assortment of products. Bloomingdales, although it also specialized in fashionable clothing, was not all that successful and certainly not at the top.

My arguments as to which firms ought to be innovators and which imitators are rather superficial, because I am not acquainted with the current economic literature on the subject. Following the tradition pioneered by Schumpeter's ground-breaking ideas on innovation and imitation, economic science has produced a large body of in-depth analysis on the subject, including mathematical models and sophisticated methods of their analysis (Iwai, 1984).

Perhaps, these considerations cast a shadow of a doubt upon Drucker's unequivocal verdict regarding the wrong strategy assumed by Macy's leaders in trying to preserve the store's traditional specialization. Perhaps, they realized that expanding the product line by introducing such alien goods as appliances would detract their attention from the store's principal item and feared they would have to relinquish their top spot in a highly competitive fashion industry. After other stores gained valuable knowledge selling a wider array of diverse products, Macy's utilized their experience to expand its own assortment.

NOTES

1. The case has not yet been published, but Ackoff was kind enough to relay the story to me and have it recorded on a cassette. For obvious reasons I have changed the proper names: Company A and Company B are oligopolistic companies in the field; Company C is the major company under consideration; Companies D, E, and L are merging with Company C; Company D-D and Company L-L are the main companies to whom the corresponding merging companies belong; the companies are located in various countries.

2. I remember my surprise when Company C's new president and CEO, "deserting" his sick company, came to our department from afar to give a talk on the theory of planning. It was, as it turned out, time well spent.

3. In essence what we have here, in terms of chess, is the problem of positional sacrifice. To the reader not familiar enough with chess, please recall that positional sacrifice is differentiated from combinational in that, at the moment of sacrifice, it is not clear how the loss will be concretely compensated, that is, how the positional advantage gained at the cost of the sacrifice will transfer into a possible effective combination in the indeterminate future.

17

Firm IV: Mechanisms Implementing the Potentiating Method

IMPLEMENTING THE POTENTIATING METHOD WITHIN THE EXISTING STRUCTURE OF A FIRM

Vertical Integration of Indicators of Firm Performance

The most obvious implementation of the potentiating method within the framework of vertical mechanisms is carried out by firms owned by a single person or a family; another area using this method is companies headed by an entrepreneur empowered by the firm's founders to govern their assets. Under these conditions, the leader's subjective vision predicated upon the potentiating method can be largely realized.

Implementing the potentiating method in a large corporation owned by many stockholders is much more complicated. Here, the firm's leaders are asked to objectify their decisions. Therefore, even if they do employ the potentiating method, they do so in a roundabout way; their decisions are veiled in the seemingly objective considerations supposed to confirm their objective nature.

Direct and explicit practice of the potentiating method is very limited in large corporations. Interesting in this connection are the views of John Kenneth Galbraith (1967). Galbraith uses Baumol's basic construction of the relation of sales and profit as the leading criterion for corporate decision making and then adds to it a few other indicators.

Galbraith has established crude ordinal priorities over the variety of goals pursued by the corporation. He correspondingly introduces a concept of primary ("a secure level of earnings and a maximum rate of growth consistent with the provision of revenues for the requisite investment"), secondary ("technological virtuosity and a rising dividend rate"), and additional ("building a better community, improved education") goals (pp. 176–77). Using my terminology,

primary and secondary goals represent material parameters, and additional goals can be assigned both material and relational parameters. The technostructure makes centralized decisions that set these priorities formally and nonformally.

Horizontal Integration of Individual Indicators of Firm Performance

With the mechanism of operation governed by a leader of a hierarchical structure, the method of potentiating can be horizontally implemented through the firm's various subdivisions. However, a number of prerequisites must be met. First of all, the given subdivisions must possess their own interests manifest in the various material and relational parameters. At the same time, the said subdivisions must be powerful enough to be able to force their interests upon the leadership of the corporation, that is, to outpower the leadership turning into a puppet manipulated by the heads of the subdivisions.

Let me illustrate this kind of arrangement. Raising the issue of profit's insufficiency for evaluating firm performance, William Baumol (1967) is skeptical of the possibility of finding one single criterion for evaluating a major firm's long-run performance. His point is that such firms have "many centers of power of varying potency and where goals have never been discussed explicitly or formulated" (p. 52).

This view was for Baumol only ancillary to his basic idea, which was based on finding a single integrative criterion. At the same time, other scholars have adopted this view as the basis for their conception of firm behavior. They maintain that the indicators that reflect the firm's performance are formed not at the top, but spontaneously through the power struggle between the firm's subdivisions. Here, the set of guiding indicators might correspond to the potentiating method; that is, along with profit, it might include the various relational parameters. This represents a peculiar case of the method of potentiating because the material and the relational parameters incorporated as independent variables arise as a result of the participants' spontaneous actions. Under these circumstances, the term "method of potentiating" should be abandoned in favor of the term "quasipotentiating." The former method (in this context) deviates from the firm's ultimate target of profit maximization because of the difficulty of coordinating the activities of the firm's subdivisions rather than because of the intrinsic nature of the firm's operation under the conditions of indeterminism. Including this rendition of the potentiating method serves to underscore the act that the vertical mechanism governing a hierarchically structured managerial system might degenerate into a horizontal mechanism implemented spontaneously through the potentiating method.

Particularly noteworthy in this connection are the views of Richard Cyert and James March — economists at the business school at Carnegie-Mellon University — expounded in their book (1963). The authors sought to unite organization theory with firm behavior theory (p. 16). According to them, a firm's organization may be seen as a coalition of different interests, representing managers, workers, stockholders, consumers, and so forth, as well as their sub-groupings, such as workers within a particular department. These interests are expressed in the differing goals that coalition members pursue.

They enumerate five kinds of goals for the firm as most important: specific requirements for production, inventory, sales, market share, and profit (pp. 41–43). They note that in some organizations there may be other goals as well, such as government orders, prestige, or tradition (p. 40). However, these do not usually play an essential role. Examining a significant number of other goals is inexpedient because, beyond the limits of the five principle goals, their value rapidly approaches the point of diminishing utility (p. 43).

Let us take a look at these objectives in somewhat more detail. First of all it can be noted that three of them, those relating to production, sales, and inventories, represent individual components of an input-output vector in real or monetary terms, on the basis of which profit — essential parameter — would have to be directly determined. At the same time, the authors see the last one as an independent goal together with the other four. Market share occupies a special place. It is a relational parameter not directly included in the input-output vector consisting of material, energy, and information goods and resources.

One of the goals, concerning production, is represented in a somewhat more complex manner than simply a material parameter because the authors believe production is actually an aggregate combination of goals. They distinguish two components or two goals in this aggregate. The first component is actually a relational parameter — smoothness. This gives me cause to think that the authors attribute market share and production smoothness to the ranks of relational parameters. The second component concerns the level of output as such, which must equal or exceed a certain value (this applies equally well to the volume of sales). Additional constraints, established by the firm's employees inasmuch as setting production volume in reaction to prices could bring the firm out of equilibrium, are explicit in this requirement. These considerations resemble the described situations that give birth to quasi-potentiating where, in order to expedite the process of convergence of an iterative procedure, profit is either supplemented with other indicators or the various constituent components of profit are set separately.

Analogous considerations enter in establishing the size of the inventories. Because different factors, difficult to define quantitatively (such as completeness and promptness in supplying the materials), influence the determination of inventory size, an absolute limit is established either on inventories or on the range of their potential worth. The value of profit in this situation is established at a desirable level, that is, sufficiently independent of the value of the other goals.

Behind each of the five goals stand the interests of the relevant coalition members, including company departments. Thus, it is obvious that the goal of production is most closely connected with the interests of the firm's manufacturing groups. Sales personnel and consumers exert the greatest influence over determining the size of inventories. The marketing department aims at increasing the firm's relative market share, and the financial sector strives for profit.

An intricate combination of the firm's many diverse objectives arises when the interests of different coalition members collide. Cyert and March are keenly aware that this sort of process leads inevitably to inconsistent decision making. How can organizations survive when irrational goals coexist in them? The problem is, as the authors note, that "the organization can remain viable by attending to the demands sequentially" (p. 35) and in a "decentralization of decision making" (p. 43).

MULTIDIMENSIONAL ORGANIZATION OF A FIRM

It thus becomes apparent that the current structure of the firm is ill-adapted for sequentially implementing the potentiating method. Granted, the method can be used in firms with a strong private leadership; but it still seems feasible to devise an organizational structure of large firms, particularly those of the corporate type, that is conducive to the comprehensive implementation of the potentiating method.

As far as I know, most attempts to resolve the organizational problems of U.S. corporations and to choose the appropriate type of leader consist of trying to modify the corporation's feudal structure. Proposed variants assume either the form of a "strong monarchy with an enlightened monarch at the top" or a system of "independent knights in charge of their own dominions." Extending this analogy to corporate leadership, we see attempts to find strong, highly qualified managers to perform the functions of the chairman of the board of directors, the president, and the senior executive officer. Otherwise, it means a return to the blessed time when entrepreneurs could execute at will their innovative projects using their own capital.

I am not familiar with any corporation attempting to create an internally developed political system (where presently operating lines of control would represent a particular case) and institute new types of leaders along with the existing ones. The profound but rather isolated suggestions concerning the governing system of a corporation seem to be very limited (Bowman, 1988).

I picked the United States as the prototype for potential reorganization of the firm along the proposed lines. I made this choice because the discrepancy between the political system of government and that of the corporations is especially pronounced in the United States. One of the key reasons for U.S. corporations losing their competitive edge in world markets seems to be the dissonance between the political system of the United States, based primarily on indeterministic ideas, and the political system of the corporations, rooted primarily in deterministic thinking. The political system in the United States represents a complex structure, which forms the so-called democratic society, but the political system of U.S. corporations, with their lack of elaborate political structures, represents a regime typical of the feudal order.

A reader may reasonably argue that a comparison between the political organization of a large system, which is descriptive of society as a whole and the political structure of a much smaller scale system such as a corporation, is rather farfetched. The argument makes sense if we take the problem to an extreme and contrast society with a firm represented by an individual artisan. However, with growth and development, the firm's organizational structure undergoes many phases gaining in complexity. Eventually, the firm reaches a stage where the characteristics of the whole — in our case, the society — tend to reverberate in the whole's constituent parts, represented here by the firms.

It seems to me that equating the structure of a large corporation with that of a developed society is reasonable in this context if only because the complexity of decision making in a large corporation with many participants is comparable with that which occurs in a small country. Having made this preamble, I want to elaborate my thesis of the inadequacy of the present day political structure of a firm for society.

In terms of political systems, a feudal order is, roughly speaking, based on the principle of a hierarchy headed by a leader towering above everyone else, even if he is partially limited by other participants. Here, the power is concentrated in a single institution. In an extreme case it is theocracy. Feudal order lacks pluralistic institutions, separation of power, or sufficient openness. Subordinates are strongly tied to the system (if not legally, then out of economic fear) and must remain loyal to the leader, being saved by him directly or via his subordinates. Feudal orders differed in many respects, such as the centralization of royal power (my vassal's vassal is also my vassal), limitations of the king (whether the king was first among equals or an unlimited monarch, possibly even a tyrant), and the degree of independence of the social status of the subordinate from the will of the monarch (aristocracy, especially inherited, possessed certain independence as a result of its social status, being largely determined not by the current leader or the position to which the aristocrat was appointed by the leader).

Apparently, the growing turbulence of the environment, rather well handled by the U.S. political system, is incompatible with the feudal organization of the firm's political infrastructure. To clarify my statement regarding the feudal structure of U.S. corporations, I want to present a system where the populace participates in all the decisions (an opposite of the feudal political scheme).

This structure, the so-called democratic society, is really a multidimensional entity. This point is not to be slighted, for all too frequently the various concepts of democracy avow some one aspect of the system — usually either the need to limit the authority of (and this is a top concern) the fact that the entire populace must take part in the decision-making process.

Modern theories of mass participation are based on a far more refined understanding of the techniques of management. Slogans expressing populist concerns of management by the workers lag far behind these modern theories that have evolved to a sophisticated understanding of the governing process, stressing participative democracy as opposed to representative democracy. Many schemes advocating participative democracy consider the actual qualifications of the various groups of workers to determine their competence to participate in the managerial process. This prevents ochlocracy as well as purely formal self-management of the firm by the workers.

Interesting in this respect is the "lowerarchic" scheme of corporate management elaborated by Russell Ackoff (1981). According to this model, workers participate in the decision-making process at various levels as determined by their qualifications. Unlike a hierarchical system, a lowerarchy assigns the leading role to the lowest level of management. With higher echelons of corporate management still remaining and decision making along this multilevel system of control being iterative, the professional holistic vision of overall development of the corporation is ensured. As pertinent as the said approach to worker participation is, it represents just one of many no less vital aspects of the problem.

The multidimensional approach considerably complicates the problem of structuring the operating mechanism of the system for all these different aspects, which may weigh differently under different circumstances and must now be integrated. I cannot withstand the temptation to remind the reader of how simple and convenient it is to describe a system (especially a complex one) based on a single dominant characteristic and then treat other characteristics as dependent variables. This approach avoids the complicated task of finding appropriate weights for various independent parameters. Therefore, the pursuit of such a single-sided approach, which captures just one (or a group) of the aspects comprising the so-called democratic system, is not accidental.

What is called democracy seems to be a multidimensional, nonlinear entity whose level of development is a function of many independent variables capable of assuming a range of different values. To emphasize this aspect, I prefer to call democracy a multidimensional society (as opposed to a nondemocratic society which, in the extreme case, is one-dimensional).

What is this multidimensional society? With respect to political systems, it incorporates such aspects as pluralism (not to be confused with relativism), democracy proper (unlike ochlocracy, power of the mob), separation of power, and openness. Let us take a closer look at components of a multidimensional society.

The pluralistic aspect in the general case assumes elaboration by independent participants of a manifold of programs having undifferentiated value, selection of a particular program (or a group of programs) for a given period based on the established procedure of priority setting, supervision over implementing the selected program, and replacement of the selected program if it proves inadequate or if the conditions change.

The democratic dimension as such reflects the number of participants involved in the mechanism of decision making and their competence and responsibility (the ability and the desire on the part of the people to assume responsibility for their actions rather than to count on someone else to provide for them and be willing to serve that someone — for otherwise how can that someone give). This forms the bases for the selection procedure to determine the participants of various institutions incorporated into the decison-making mechanism.

Separation of power in a multidimensional society in turn entails many aspects. Of fundamental importance is institutionalization of more or less equally powerful branches, both with respect to the scope of their activity (especially the separation of the ideological or religious function) and their territorial (central power or regional) and functional responsibilities. For the system as a whole, the last feature manifests itself in the separation of power between the bodies representing national unity or oneness and national traditions (for instance, a constitutional monarch of the Anglo-Saxon variety) and other bodies that fulfill specific functions. The latter include (as part of the system of checks and balances) the legislative bodies, which lay down laws of fundamental long-range significance (constitution, for instance) and non-conflicting, more specific laws that are reviewed more often (parliamentary laws); the executive power, which carries out these laws according to the situation at hand; and the judicial branch, which sees that the laws, as well as the laws for changing the laws, are abided by properly.

Finally, openness refers to immigration and tourism into the country, penetration of new ideas and imports as well as emigration and tourism of people, ideas, and goods out of the country.

I believe the diversity of pluralistic mechanisms described in Chapter 3 is the heart of a multidimensional society. I do not mean to negate other independent dimensions but merely wish to emphasize the importance of the pluralistic one. Particularly interesting for corporations is a type of pluralistic mechanism that allows for disequilibrium to arise — because of the creation of a manifold — and then transforms this manifold into a singular variety by overseeing its implementation and replacement. Naturally, progress in other areas of an inte-grative society is conducive to the perfection of the pluralistic mechanism itself. At the same time, their development is important in itself, for society can be perfected with the pluralistic dimension kept intact.

In creating a multidimensional society, one has to be extremely cautious because of the complex system of priorities of each individual aspect. For this reason, disagreements between political leaders often revolve not so much around the ideal (which many people belonging to a given culture may aspire to) but around ways of pursuing it and the speed of change along the different dimensions. Here, radicals stand out in that they advocate drastic simultaneous changes along all dimensions. In many societies, a radical approach calling for an all-round transformation would in all probability lead to progress along some dimensions but cause regress in others, eventually driving the system toward one-dimensionality.

The leadership structure of a corporation exhibits all the aspects of social mechanisms mentioned above in their embryonic stage. The real power in the firm resides with the executive branch (Mintzberg, 1983).

In the following discussion I want to focus on one of the primary issues of restructuring firm management, the creation of pluralistic mechanisms. In passing I will touch upon other aspects of the mechanism of management.

Pluralistic mechanisms represent the core of the firm's managerial structure. They form the foundation upon which the diversity of the courses of development can be formed. In order for each alternative plan to unfold and especially assume definite shape in the form of a specific strategy of action, it is essential that each strategy be developed by an independent unit. Each unit must concentrate its effort upon its own program and, in addition, possess the necessary resources: qualified individuals, money, and time.

Moreover, nobody from the outside should impose any individual preferences.

Who is in charge of developing and approving strategies in the corporation today? These tasks are carried out via appointed personnel; in other words, firms lack independent internal organizations empowered to elaborate new avenues of development. Commonly the development of alternative programs is assigned to outside organizations, such as consulting agencies because employees of the firm immersed in its current affairs cannot devote their time to developing new strategies. In itself, collaboration with consulting groups is useful although possibly insufficient. These consulting agencies are hired by the firm's leadership and must, therefore, bow to the interests of their clients. Inevitably, this leads to conformity on the part of the consulting agencies although the degree may vary considerably.

Moreover, development of corporate strategy by outside organizations without the participation of the firm's leadership creates serious problems when it comes to actual realization of the proposed strategy. In elaborating new courses of development, it is important to involve personnel who will be responsible for their implementation.

A certain ineffectiveness of relying upon outside organizations to devise new avenues of development gave rise to various theories of employee participation (in collaboration with consulting agencies and other interested parties). A few years ago we witnessed the emergence of so-called team planning (Gross, 1986). The task of these teams is to devise significantly unorthodox alternatives that circumvent planning horizons set by the leading staff. These teams composed of the leading representatives, with different backgrounds and from various departments, undertake a joint effort to put together an innovative and realistic plan of development. Sometimes the operation is limited to one such team, and sometimes several competing teams work simultaneously.

Whatever the worth of these groups, diverting management's attention from current affairs to long-term projects is objectionable, in spite of the seeming attractiveness of such schemes (a single individual embodying both developer and executive functions). And these teams are appointed; they are comprised of employees subordinate to one governing body. Inevitably, they must accommodate the interests of the firm's leadership. Of course, the act that they do not all conform to the same degree cannot be ignored.

All this leads to the conclusion that corporations today lack independent bodies to form strategic policies. As a result, a single well-designed program gets passed for approval to the legislative body (the board of directors) and then to the supreme authority in a corporation, the annual meeting of stockholders. The board of directors, consisting mostly of outsiders — people whose interest in the affairs of the firm is of a rather episodic nature and who lack their own strong body of people — is not really in a position to propose alternative plans. At best it may perform a critical overview of the program or reject it. Stockholders' options are subject to the same limitations. They can criticize or reject the proposed plan with no viable alternatives to offer because a new plan requires long-term collective commitment, unmanageable by an isolated "atomized" group of stockholders.

So in reality businesses are devoid of any separation of power. The board of directors and the stockholders' annual meetings are but a feeble imitation of real legislative power. Not surprisingly, the executive branch can manipulate both the board of directors (even interfering in panel selection) and the stockholders.

Aforementioned deficiencies call for creating within the corporation independent institutions that, like the political parties, support themselves through the participants: they are not appointed, confirmed, or hired to serve. These corporate parties will be involved in elaborating alternative courses of development. These programs, representing pure strategies, must be subsequently distilled into mixed strategies reflecting the diversity of strategies proposed by independent parties.[1]

Program development must be accompanied by the separation of power, legislative power in particular. Actual corporate operations will be based on a mixed strategy. In addition, parties must include institutions that oversee proper execution of the selected program. Finally, there must be institutions to revise or retrieve the acting program if it proves inadequate or if conditions change. The cycle repeats by reverting to the manifold of programs — pure strategies, subsequent elaboration of a mixed strategy, and so on.

The principal question that arises in connection with the firm's management is the selection of the participants in the pluralistic mechanism. They must be responsible and competent individuals motivated by the success of the corporation.

It seems to me that when it comes to long-range issues, the decisions should rest with the stockholders. Stockholders have proved to be people responsible for their future if only because they have agreed to divide their income between consumption and savings. Presumably (and this is only true statistically) people who consume all their income, unless they are starving, of course, are less responsible. Furthermore, stockholders did more than just choose to save, something they could do at a bank where their assets are insured. They decided to invest their savings in certain businesses. To undertake such a step requires competency and responsibility.

Being a stockholder is probably not sufficient qualification to be a participant in a strategic decision-making mechanism. Among stockholders there are people and organizations (especially pension funds) who would sacrifice the firm's long-range interests in favor of current gain. Presumably, the family of stockholders as a whole is capable of overcoming this impatience of the heart.

Admitting the firm's employees, even if they are owners (who bought the firm), into the pluralistic mechanism is objectionable for the following reason. These employees, who pursue primarily short-term interests, may ignore long-range concerns such as capital investments in future development, and their purchase of the business could be a result of a desperate step taken at a moment of crisis (such as bankruptcy or liquidation). This does not exclude employees from participating in the mechanism of operation but limits them to the framework established by the stockholders.

Putting this multidimensional mechanism into the center of the firm's managerial practices (a moderate constraint upon the leader) means that movement away from the center will be represented by either strong or weak limitations. The polar case of strong limitations is a totally bureaucratized system of operations based on a predetermined criterion of greatest welfare. Weak constraints reflect greater discretionary powers of the leader, in the extreme case a tyrant.

TYPOLOGY OF FIRM LEADERSHIP

Of all the different types of corporate leaders, the most representative ones today are the manager and the entrepreneur. When it comes to innovation, the entrepreneurial edge hardly needs any proof. In principle, an entrepreneur is free to proclaim his idea, being spared the burden of having to prove its soundness to anyone. At the same time he is not limited in its implementation by any internal constraints of the organization, putting primarily his own capital on the line.

A manager is a hired hand who administers other people's capital. Therefore he is required to show the validity of all his major decisions and is given rather limited rights in directing the corporation. These circumstances curb the manager in proposing innovative ventures unless he is able to substantiate them strongly. Moreover, not only is the manager unable to propose risky ventures, but he would not be granted sufficient leeway if their implementation required too much freedom of action on his part.

Therefore, a managerial position is best adapted to handle situations calling for the implementation of relatively risk-free innovations or the supervision of ongoing newly introduced technology even if it does entail modification. Hence, there is the dream of introducing innovations through the entrepreneur. Nevertheless, the

scale of present-day production is so large and capital requirements are so great that a pervasive emergence of leaders-entrepreneurs is impossible.

Specific types of leaders are needed to implement the diverse methods of managing a firm. My basic premise in constructing leadership typology is to have a variety of different kinds of leaders whose adequacy for a given organization is determined by the requirements of that organization and the types of tasks it performs. The existing diversity in leadership should be preserved and expanded by introducing new, hereto unknown types. This formulation of the problem naturally leads us to derive the leadership types using a deductive scheme. Current research into this typology is rather atomized and descriptive in nature.

I want to represent this deductive approach through a multidimensional matrix. Similar to the periodic table of elements, the deductive scheme represented in the form of a matrix reveals both the existing types of leaders (administrators and managers) and the new ones (navigators, for instance). This may help us to better understand the types of leaders that companies lack in order to deal with the ever increasing complexity of developing and implementing innovative ideas.

Tackling the problem in its entirety is, of course, too formidable a task. I shall examine two assumptions crucial to outlining this deductive scheme. They concern the leader's responsibilities (in a broad sense of the word, that is, duties, obligations, and liabilities) and the rights (understood broadly to include freedom, privileges, and independence). These assumptions are limited to internal rules of an organization; external constraints imposed by the federal or local governments are omitted from the scheme.

In examining leadership behavior it may appear that the leader's responsibilities are inversely related to his rights: the more rights he has, the fewer his obligations. In an extreme case, the leader possesses all the rights and has no obligations. Nevertheless, rights and responsibilities represent two independent characteristics, which may seem quasi-opposite on the surface but are really complementary to each other.

In turn, measure is introduced on each of the two dimensions by punctuating them with phases. As a first approximation, I distinguish three phases and the corresponding methods of operations. My choice of three phases, just like any other division into three parts, avoids reducing the respective parameter to two extreme states so that a distinction can be made between extreme and intermediate states while eschewing all the complications of analysis over a continuum. I realize that this division may prove inadequate in describing the intricate variety of conditions within the firm, but it may serve as a good first approximation.

Naturally, a number of conditions mentioned above as well as some others not taken into account detract considerably from my analysis. Still, the discussion is worthwhile, for these conclusions may stand even under an expanded set of conditions. Following a systems approach, the assumptions I have made in constructing leadership typology can be analyzed from the functional, the structural, the processional, and the genesis points of view.

From the functional point of view, the following picture emerges. The first assumption concerning the responsibilities of the leader pertains to the degree to which the development of the organization (system) pursues radical changes in the environment. It reflects how innovative the proposed plans are. Thus arise the three categories of pioneering ideas (in the sense of radically new ones), significantly new ideas, and modifications of the known ones.

The functional role of the second assumption is the leader's right to determine the extent of the limitations imposed by internal rules. Here, cases are characterized as gravitating toward the leader totally unrestrained by the internal rules of the corporation, intermediate cases where existing rules are flexible enough to accommodate the leader in using an appropriate method in his decision-making process, and complete restriction of the decision maker by the rules in question.

From the structural point of view, the first assumption reflects the degree of completeness and consistency of the relationship between a given organization and its environment (its suprasystem). Here, at least three types of interface are possible: lack of a complete and consistent relationship, partial incompleteness and inconsistency, and complete and consistent relationship. This classification also applies to the second assumption concerning the internal structure of corporate leadership.

From the processional (operational) point of view, our first assumption pertains to the basis of decision making: directing, potentiating, or programming. The second assumption reflects the presence of a leader who is an unlimited monarch, limited by the course of the firm's development, or limited by the plan (program) of the firm's development.

From the genesis point of view, the first assumption is manifest in the evolution of leadership styles — from entrepreneurs to various types of managers and possibly to new types of leaders, such as navigators. The historic significance of the second assumption lies in the development of the internal structure of corporate management beginning, in political terms, with absolute monarchy, moving to limited feudal structures, with a subsequent transition to a multidimensional society.

The two assumptions underlying the deductive scheme of the typology of firm leadership are the objectivity of the leader's decisions and the limitations imposed upon him by the internal rules of the organization. Cases displaying complete independence of the two assumptions usually involve leaders who are unlimited (have weak limitations), no matter what procedures of verification are used. Cases where limitations imposed upon the leader by strong or moderate internal rules of the organization will exhibit a strong correlation with the structure of these rules. Rule structure can be distinguished by the restrictiveness of these rules in the application of past results to the future — this is called conditions — and also by the degree to which today's results will affect the leader's future activity — this is called results.

My use of traditional terminology in characterizing both the results and the conditions prompts me to clarify their usage. Here, the difference between a plan and a program consists in the degree of detail specified in a plan. By platform I mean the overall course of development that represents pure strategy. This does not require a pluralistic mechanism in its entirety, only the part dealing with creating the manifold. Transforming a manifold into a variety is nothing but applying a code of laws used to transform a number of pure strategies into one mixed strategy.

Table 17.1 shows correlations between required objectivity in the leader's decisions and the structure of the internally imposed rules.

Table 17.2 represents the typology of leadership.

The extreme case (33) corresponds to a type of leader who could be called an administrator. This term is usually associated with a leader whose decisions are amenable to complete objectification and who, at the same time, acts strictly in accordance with the rules of the organization. Any discretionary power presupposing less rigid rules, as well as incomplete objectification of the decisions, are all secondary in the work of an administrator.

TABLE 17.1
Correlation between Objectivity and Limitations

	Limitations Imposed upon the Leader		
Objectivity of Decisions	*Conditions/ Results*	*Moderate*	*Strong*
Subjective	Conditions	Contract	Dogma
	Results	Sketch	"Encyclics"
Semi-subjective	Conditions	Pluralistic mechanism	Code of laws
	Results	Platform	Mixed strategy
Objective	Conditions	Basic rules	Manual
	Results	Plan	Program

TABLE 17.2
Typology of Corporate Leadership

	Limitations Imposed upon the Leader		
	Weak	*Moderate*	*Strong*
Subjective	11	12	13
Semi-subjective	21	22	23
Objective	31	32	33

The opposite case could be called a pioneer, case (11). Typically, decisions made by this type of leader are evaluated based on belief. The leader is also practically uninhibited by the internal rules of the organization. These considerations are important in understanding the role of entrepreneurship in the realm of pioneering ideas.

First of all, the pioneer has advances in the fields of new ideas whose validity is very hard to prove. What is needed here is faith. Second, any preset organizational rules curbing a pioneer, who in this case needs great flexibility, would be detrimental to developing the proposed ideas. As a rule, a pioneer must have personal capital with which to pursue a vision. Examples of such pioneers are inventors of new gadgets, for instance, designers of the Apple personal computer. A pioneer must satisfy two assumptions: faith as a basis of verifying his decisions and considerable freedom from the internal rules of the organization.

A creative manager is different from a pioneer. Although the former is involved in introducing novel ideas, they must be strongly validated, for the manager puts other people's capital at risk. Therefore a manager is limited by the rules of the organization (less than an administrator but more than a pioneer). The manager assumes slot (32) in the typology.

Examples of pioneers and managers reveal a strong correlation between the objectivity of a leader's decisions and his freedom from the internal rules of the organization. Nevertheless, these two aspects, which determine the type of leader, are to an extent independent. From the history of the church, we know of institutions where the leader's choice is accepted on faith while he is strongly limited within the organization. This applies to the Pope as the head of the Catholic Church. This type of leader represents case (13) in the matrix.[2]

An example of a contrasting situation is one where the leader's decisions are sanctioned but where there are limiting internal rules, case (31). It may be expedient, for one reason or another, especially in crisis situations to give the leader unlimited powers within the organization while subjecting the results to stringent verification.[3]

Except for the (12) slot all the cells of the upper and lower rows of the matrix have been filled. In this empty slot we can place leaders of venture capital firms. Their decisions are accepted largely on faith, but limitations imposed upon them by the internal rules of the organization are more severe than in the case of a pioneer. Here, a major difference is the method of decision verification: the direct approach is replaced by a more risky indirect one. What is evaluated is not the outcome or the idea, but the leader as an individual. This judgment can be based on any method — directing, potentiating, programming, or their combination. Therefore, if faith in the leader's decision is preserved, the emphasis shifts toward the process of selecting the leader. The choice is largely determined by the trust invested in him as a person.

The three slots belonging to the middle row remain open. It seems that (21), which represents an aesthetic method of decision making combined with weak internal limitations, corresponds to certain types of leaders of large companies whose ideas are directed at enhancing the firm's potential, at improving its position on the market. Managers combining the functions of the chairman of the board of directors, the president, and the chief executive officer resemble type (21). They differ from pioneers in that the latter are strongly goal-oriented toward such aims as introducing new types of products or implementing a new technology.

The last slot in the matrix (23), corresponding to the potentiating method of decision verification and strong internal constraints, is perhaps characteristic of such a type of leader as the coach of a sports team.

Much trickier is the case of leaders, especially in large and medium-sized companies, whose decisions are based on the potentiating approach but who lack absolute (or nearly absolute) power characteristic of (12) or are severally limited like an administrator (33). Here, leaders have considerable amounts of capital at their disposal, so it is dangerous to give them too much discretion-ary power. Limiting their role to that of an administrator, however, would be counterproductive in terms of the flexibility required to implement innovative ideas.

In view of all this, our discussion concerning the creation of a multidimensional political system within a corporation acquires a special significance. I think that a pluralistic system, in conjunction with democracy (again, not to be confused with ochlocracy), separation of power, and openness would lay the foundation for a system that would provide sufficient freedom to the type of leader whose decisions are amenable only to evaluation of a potential. Indeed, within the framework of such a system, the leader acquires a new set of conditions in the form of a platform of the winning party to which he belongs (perhaps even heads) or a modified platform —

mixed strategy — if the platform must pass through the legislative body composed of several parties. This platform provides sufficient leeway for the leader while binding him to certain guidelines instituted by competent and responsible individuals representing the firm.

The leader within such a political system (22) could be called a navigator. This underscores his role as an originator of new avenues of development aimed at increasing the firm's potential. Here perhaps the term "navigator" is more precise than the term "president" or "chairman of the board of directors," for the latter are associated with different types of leaders.

In conclusion, the introduction of just a rather limited number of dimensions (two in our case) necessary to construct the type of leader creates an extensive diversity. This diversity is vital for different situations calling for different types of leaders. At the same time, a new figure heading this typology emerges, a navigator. This category is lacking in U.S. business today. Nevertheless, a navigator is vital for U.S. firms that need to reorganize to improve their potential to create and implement pioneering technologies.

Experimental verification of the new system of corporate performance must precede this reorganization. It would be ill advised to try to force this system upon large corporations because its failure is fraught with exorbitant losses. At the same time, this system does not lend itself to implementation at small firms because they lack a sufficiently elaborate infrastructure. It seems most appropriate to gradually introduce this scheme to developing companies as they evolve.

NOTES

1. In the process of creating a mixed strategy, an ugly program (a program monster) could appear because one cannot take the best feature per se from each program: each of these best features is the best relative to a certain program. In this connection the following thought of Charles Montesquieu (1964) comes to mind.

> Western poets tell how a painter, wanting to make a portrait of the goddess of beauty, took from each of the assembled beauties of Greece her most lovely feature and combined them into a single likeness of the most beautiful of all goddesses. [Zeuxis (fifth century B.C.), a Greek painter, was reputed to have used five different models when he painted Helen from the Agrigentides.] If a man had concluded, on that account, that she was both blond and brunette, had both blue and dark eyes, or that she was gentle and haughty at the same time, he would have passed for a fool (pp. 120–21).

2. I am very grateful to Alton Long for pointing my attention to this fact.
3. For instance, these conditions prevailed in the Soviet Union at "Dalstroi," a rather remote corner of the GULAG archipelago.

Conclusion

I believe the various issues raised throughout the book put us in a position to answer, if only in part, the seven questions posed at the beginning.

1. In principle, is it possible to judge unequivocally a course of economic development based on technological progress?
2. Is there some kind of order within the chaos of the stock market?
3. Can an outsider (an expert, in particular), rather than the creator, himself, judge the advisability of financing new trends in science, even of applied science, directed toward a specific and final practical end?
4. Is it possible to derive an index of economic growth based on current prices, in particular, by taking prices for a certain period as constant prices?
5. Why is profit, expressed even in probabilistic terms, insufficient in evaluating a firm's performance? Why must profit be supplemented by other criteria characterizing the relations between the firm and its environment?
6. When a firm is faced with several alternative projects (described in probabilistic terms) for developing a new product, is parallel implementation ever warranted?
7. Can we say that an analytical representation of the consumer demand function is linked with a deterministic approach whereas an optimal model is linked with an indeterministic one?

The answers are predicted on the suggestions of various phases within the indeterministic spectrum. For instance, the directing phase, at which no single course of economic development can be proved (or refuted) to be the best, one's decision with regard to the proper course of development is grounded only in personal beliefs. The first of the seven questions posed above can be addressed from this perspective.

The possibility of uncovering certain dynamic features of the system as a whole in a chaotic environment shows light at the end of the tunnel when we address the second question.

The other five questions belong to that part of the indeterministic spectrum where the process unfolds along the selected course of development. Here, we can devise adequate, although inconsistent or incomplete methods of integrating the elements of the system. One such method is based on creating predisposition of the system to further development. This method of linking the system's elements provides clues to questions three, four, and five. Showing the validity of the said method of computing predispositions really constitutes the core of my work.

Next in line are sufficiently complete and consistent methods that yield, in terms of probabilities, a certain unique course of development. This class of method is useful in answering question six.

Finally, some methods ensure a definite complete and consistent course of development of the system. These methods range from simple reactive ones to extremely sophisticated selective methods including possible transmutations. The distinguishing feature of reactive methods is that input determines the output. Selective methods underscore the driving forces (goals, for instance) that guide the process of selection within the feasible space. The results of applying these methods are described in terms of full certainty and are invariant to the method utilized. The answer to question seven is rooted in the mutual transformation of these methods with the result completely determined.

I suggest the present work strives to advance a new approach to the analysis of the economic system, rooted in an indeterministic vision. At the heart of this approach is another phase of the indeterministic spectrum associated with the aesthetic method and its constructive counterpart, the potentiating method — calculus of predispositions. A number of immediate and rather practical results stem from the concept of potentiating, and I would like to focus on these.

In a macroeconomic analysis the use of a potentiating method allows us to solve the problem of measuring economical dynamics in a different manner, separating the search for pace of economic growth from the measurement of economical development. In those cases when there is a possibility of constructing chain indices, essentially it is a matter of measuring the growth of the economy. Otherwise, one should seek the pace of development using the technology of calculating predispositions. Considering the role of the growth index in a macroeconomic analysis, there is hardly any need to prove the importance of realizing the bounds of its utilization.

An analogous situation takes place regarding measuring various countries' development, which should be calculated through the discovery of their predispositions — potentials. The current methodology of comparing levels of development of countries — used by the United Nations, in particular, for setting the dues of member countries — is based on the levels of GNP measured in the prices of a base country. This methodology is used even though the index depends on selecting the prices of the base country. Needless to say, the calculation of levels of development through setting equivalents in the value of these countries' currencies does not change the essence of the indicated difficulties because one still must select a set of products common to all countries.

On the level of firms, the potentiating method enhances the intuition of firms' managers by providing a certain procedure for evaluating material parameters and designing paths of formation of relational parameters and their evaluations. Much of what managers have been doing will receive conceptual clarification with all the corresponding consequences.

The application of my concept of the potentiating method in the development of the planning system at Latin American Division Armco by Julio R. Bartol (vice-president, ARMCO, Inc., and president, Latin American Division) and Ali Geranmayeh (partner, The Institute for Interactive Management [INTERACT]) could serve as an illustration of the last statement. (I referred to their work in Chapter 16.)

Further comprehension of the potentiating method allows us to dismember the methods of strategic and tactical planning, keeping in mind the conventionality of these terms. The potentiating method is closer to strategic planning; the programming method is closer to tactical planning. One of the mistakes of the strategic planning departments has been that the plans they elaborated constituted a program, that is, attempts to predict the future of the firm through a complete relation of initial and future states. Programs of this sort produced by the headquarters of a firm for line managers shackled the initiative of the latter. Hence, it is no accident that some U.S. firms liquidated strategic planning departments and made planners subordinate to line managers.

Just as the potentiating method is helpful in explicitly showing the difference between the various mechanisms of management in time, it is equally pertinent in space. The director of the American-Japanese Institute (Philadelphia), Dr. Toshiaki Taga, was kind enough to tell me of a paper (Kagano, 1983) dealing with the importance of various objectives of several hundred large manufacturing firms as judged in 1979 by U.S. and Japanese managers.

The ranking system was: very desirable = 3, desirable = 2, so-so = 1, not so = 0. The lexicon of the potentiating method presented in my work can be used to construct a table summarizing this paper.

Table C.1 reveals unequivocally the principal difference between U.S. and Japanese companies. U.S. firms place much greater emphasis upon the material parameters (almost twice as much for profitability and tenfold for positive performance of company stock), and Japanese companies stress positional ones (market share, twice as much; corporate image, four times). Also worth noting is the ratio of the value of material parameters to total value of all parameters: 60.1 percent for U.S. companies and 20.9 percent for Japanese companies.

The potentiating method provides an opportunity to clearly distinguish the differences between the types of workers suitable for upper management in firms focusing upon creating predispositions and the lower-level managers, who are required to possess greater knowledge of concrete operational methods and who lean toward programming methods. At the same time, comprehension of the potentiating method by the latter allows them to better understand the requirements of upper management. It is a kind of tunnel effect, where managers at different levels move toward each other. This is particularly important when higher-level managers are involved in complex decision making, which makes it difficult for them to explain the solution they adopted, mostly through intuition, to lower-level managers in terms the latter understand.

Thus, the potentiating method allows one to formulate clearly the need for subjectivity. Indeed, on the one hand, strategic thinking demands subjectivity. On the other hand, managers (unlike private owners) are frequently forced to objectivize all major decisions. In this situation managers have to either abandon the strategic decisions or tailor their objective calculations to fit them. Whatever the case may be, it is clear that, rather than just searching for new ways to objectivize the solutions, managers must create conditions under which this method will be utilized. The importance of this remark comes to the fore with the advent of computers, which create the illusion that it is possible to obtain all-encompassing objective information, and the appearance on the market of the PIMS model (examined in Chapter 15), which claims to produce objective information. (There is hardly any need to prove the enormous role of computers and the PIMS model as means of assisting firms' managers in decision making. One leader of a major U.S. firm said after my lecture on the potentiating method that this method grants him a license for subjectivity.)

Everything said up to now also lays the methodological foundation for setting the price of a business, whether the sale of the

TABLE C.1
Comparative Evaluation of Parameters Reflecting the Performance of U.S. and Japanese Firms

	Evaluations					
	U.S.		*Japan*		*Comparative U.S. & Japan*	
Objectives	*Scores*	*%*	*Scores*	*%*	*Scores*	*Ratio*
Essentials						
Return on investment	2.43	140.91	1.24	120.60	+1.19	+1.96
Rise in stock prices	1.14	19.19	0.02	0.33	+1.12	57.00
Subtotal	3.57	60.10	1.26	20.93	+2.31	+2.83
Relations						
Market share	0.73	12.29	1.43	23.75	–0.70	–1.96
Improvement in product portfolio	0.50	8.42	0.68	11.30	–0.18	–1.36
Rationalization of production-distribution system	0.46	7.74	0.71	11.79	–0.25	–1.54
Entry/asset ratio	0.38	6.40	0.59	9.80	–0.21	–1.55
New product ratio	0.21	3.54	1.06	17.61	–0.85	–5.05
Improvement of corporate image	0.05	0.84	0.20	3.32	–0.15	–4.00
Improvement of working conditions	0.04	0.67	0.09	1.50	–0.05	–2.25
Subtotal	2.37	39.90	4.76	79.07	–2.39	–2.01
Total	5.94	100.00	6.02	100.00	–0.08	–1.01

firm's own business or the purchase of someone else's — a commodity whose role in the market is growing by leaps and bounds. It is well known that it is insufficient to set the price of a business on the basis of the balance sheet of the firm's capital or the price of its stocks. The holistic effect of purchasing a firm can be grasped through the potentiating method.

Finally, understanding the potentiating method may prompt a restructuring of the system of business education. Currently the business schools' applicant selection process puts a tremendous focus on knowledge of mathematical methods requiring, for the most part, development of the brain's left hemisphere. Meanwhile, to a great extent the thinking of firm managers requires development of the right hemisphere, which seems to be predominantly responsible for the potentiating method. Eliminating applicants with a more developed right hemisphere may deprive firms of potentially gifted managers. I find worthy of discussion the issue of relevance of maintaining at business schools two groups with different curricula depending on the students' predisposition to the potentiating or programming method of thinking.

New problems of understanding the economic mechanisms operating under the conditions of indeterminism arise in this book. Issues raised here are outside the scope of orthodox trends in economics, which are based primarily upon deterministic and probabilistic notions and respective models of equilibrium. Naturally, what I have said does not negate all the contributions of deterministic and probabilistic economic analysis; it merely limits the scope of applicability.

To quote an outstanding Russian literary critic Michail Bakhtin (1984) writing about the poetics of Dostoyevsky:

> Thus no new artistic genre ever nullifies or replaces old ones. But at the same time each fundamentally and significantly new genre, once it arrives, exerts influence on the entire circle of old genres: the new genre makes the old ones, so to speak, more conscious; it forces them to better perceive their own possibilities and boundaries, that is, to overcome their own *naivete*. Such, for example, was the influence of the novel as a new genre on all the old literary genres: on the novella, the narrative poem, the drama, the lyric. Moreover, a new genre can have a positive influence on old genres, to the extent, of course, that their generic natures permit it; thus, for example, one can speak of a certain "novelization" of old genres in the epoch of the novel's flowering. The effect of new genres on old ones in most cases promotes their renewal and enrichment (p. 271).

Bibliography

Ackoff, R. *A Concept of Corporate Planning*. New York: Wiley-Interscience, 1970.

____. *Creating the Corporate Future*. New York: Wiley, 1981.

Ackoff, R., Broholm, P., & Snow, R. *Revitalizing Western Economies*. San Francisco: Jossey-Bass, 1984.

Ackoff, R., & Deane, R. "The Revitalization of ALCOA Tennessee Operations." *National Productivity Review*, Summer 1984, pp. 239–45.

Ackoff, R., & Emery, F. *On Purposeful Systems*. Chicago: Aldine, 1972.

Ackoff, R., & Emshoff, J. "Advertising Research at Anheuser-Busch, Inc." *Sloan Management Review*, Winter 1975, pp. 1–15.

Adams, J. *Conceptual Blockbusting: A Guide to Better Ideas*. New York: W. W. Norton, 1980.

Arrow, K. *Social Choice and Individual Value*. New Haven: Yale University Press, 1963.

Arrow, K., & Hurwicz, L., eds. *Studies in Resource Allocation Processes*. New York: Cambridge University Press, 1977.

Ashby, W. *An Introduction to Cybernetics*. London: Chapman & Hall, 1956.

Atkin, R. "From Cohomology in Physics to q-connectivity in Social Sciences." *International Journal of Man-Machine Studies* 4 (1972): 139–67.

Bakhtin, M. *Problems of Dostoyevsky's Poetics*. Minneapolis: University of Minnesota Press, 1984.

Barry, D., & Fristedt, B. *Sequential Allocation of Experiments*. London: Chapman & Hall, 1985.

Bartol, J., & Geranmayeh, A. *Strategic Planning at Armco's Latin American Division*. Philadelphia: Interact, 1989.

Baumol, W. *Business Behavior, Value, and Growth*. New York: Macmillan, 1959.

____. *Business Behavior, Value, and Growth*. New York: Harcourt, Brace, & World, 1967.

Becker, A. *The Burden of Soviet Defence*. Santa Monica: Rand, 1981.

Bellman, R. *Dynamic Programming*. Princeton: Princeton University Press, 1957.

Bergson, A. *The Real National Income of Soviet Russia*. Cambridge: Harvard University Press, 1961.

Bergson, H. *Creative Evolution*. New York: Holt, 1911.

Bertalanffy, L. *General System Theory*. New York: George Brasilier, 1968.

Birkhoff, G. *Aesthetic Measure*. Cambridge: Harvard University Press, 1933.

Boulding, K. *Beyond Economics*. Ann Arbor: University of Michigan Press, 1968.

____. *Economics as Science*. New York: McGraw-Hill, 1970.

____. *A Preface to Grant Economics: The Economy of Love and Fear*. New York: Praeger, 1981.

Bowman, E. *Strategy Changes*. A working paper of the Reginald H. Jones Center, Wharton School, University of Pennsylvania, 1988.

Buzzel, R., & Chussil, M. *The Full Potential of Your Business*. Cambridge: SIP, 1983.

Buzzell, R., & Gale, B. *The PIMS Principles*. New York: Free Press, 1987.

Coffey, W. *303 of the World's Worst Predictions*. New York: Tribeca Communications, 1983.

Copeland, T., Koller, T., & Murrin, J. *Valuation: Measuring and Managing the Value of Companies*. New York: Wiley, 1990.

Cyert, R., & March, J. *A Behavioral Theory of the Firm*. Englewood Cliffs, N.J.: Prentice Hall, 1963.

Danilov-Danil'ian, V., & Zavelsky, M. *The System of Optimal Strategic Planning of an Economy*. Moscow: Nauka, 1975.

Dantzig, G. *Linear Programming and Extension*. Princeton: Princeton University Press, 1963.

Dennis, J. *Mathematical Programming and Electrical Networks*. Cambridge: MIT Press, 1959.

Domotor, Z. *Non-deterministic Probability*. Manuscript, University of Pennsylvania, 1976.

Drucker, P. *The Practice of Management*. New York: Harper & Brothers, 1954.

———. *Innovation and Entrepreneurship*. New York: Harper & Row, 1985.

———. "Reckoning with the Pension Fund Revolution." *Harvard Business Review*, March–April 1991, pp. 106–14.

Dudkin, L., Rabinovich, I., & Vakhutinsky, I. *Iterative Aggregation Theory*. New York: Marcel Dekker, 1987.

Dynkin, E. "Some Probability Models for a Developing Economy." *Soviet Mathematical Doklady* 12 (1971): 1422–25.

———. "Kolmogorov and the Theory of Markov Processes." *The Annals of Probability* 17 (1989): 822–32.

"Economic Systems." *Encyclopedia Britannica*, Macropedia, Vol. 6, 1983.

Einstein, A. *Relativity*. New York: Crown, 1961.

Encyclopedia Americana, Vol. 9, 1986.

The Encyclopedia of Philosophy, Vols. 1 & 2. New York: Macmillan, 1972.

Ershov, E. "On Finding and Utilizing the Structural Properties of Matrices in Planning Problems." *Economics and Mathematical Methods* 2 (1966): 249–61.

Eysenck, H., & Castle, M. "Training in Art as a Factor in the Determination of Preference Judgments for Polygons." *British Journal of Psychology* 61 (1970): 65–81.

Farjoun, E., & Machover, M. *Laws of Chaos: A Probabilistic Approach to Political Economy*. London: Verso, 1983.

Fedoseev, A. *The Trap*. Frankfurt-am-Main: Posev, 1976.

Feynman, R. *The Character of Physical Law*. Cambridge: MIT Press, 1965.

Fine, D. "Mineral Resource Dependency Crisis." In *The Resource War in 3-D*. World Affairs Council of Pittsburgh, 1980.

Galbraith, J. *The New Industrial State*. Boston: Houghton Mifflin, 1967.

Gale, B. "The PIMS Competitive Strategy Paradigm." SPI Membership Conference 1986. Cambridge: The Strategic Planning Institute, 1986.

Georgescu-Roegen, N. *The Entropy Law and the Economic Process*. Cambridge: Harvard University Press, 1971.

Gerschenkron, A. *A Dollar Index of Soviet Machinery Output, 1927–28 to 1937*. Santa Monica: RAND, 1951.

Gharajedaghi, J. *Toward a Systems Theory of Organization*. Seaside, Calif.: Intersystems, 1985.

———. *A Prologue to National Development Planning*. New York: Greenwood Press, 1986.

Gittens, T. *Multiarmed Bandit Allocation Indices*, 1989.

Gleick, J. *Chaos*. New York: Viking, 1987.

Goldberg, D. *Genetic Algorithms in Search, Optimization, and Machine Learning*. New York: Addison-Wesley, 1988.

Gomory, R., & Schmitt, R. "Science and Product." *Science* 240 (1988): 1131–32, 1203.

Gordon, R. "Short-Period Price Determination." *American Economic Review* 38 (1948): 265–88.

Green, D., & Higgins, C. *SOVMOD I: A Macroeconometric Model of the Soviet Union*. New York: Academic Press, 1977.

Griliches, Z. "R&D and Productivity: Measurement Issues and Econometric Results." *Science* 237 (1987): 31–35.

———, ed. *Price Indexes and Quality Change*. Cambridge: Harvard University Press, 1971.

Gross, N. "Corporate Revitalization via Team Planning." In *Handbook of Business Strategy, 1985/1986 Yearbook*. Boston: WGL, 1986.

Hegel, G. *Aesthetics*. Moscow, 1968.

Heinrich, B. *Bumblee Economics*. Cambridge: Harvard University Press, 1979.

Himmelblau,D. ed. *Decomposition of Large-Scale Problems*. New York: North-Holland, 1973.

Hingers, R., & Willer, D. "Prevailing Postulates of Social Exchange Theory." In *Theoretical Perspectives in Sociology*. Ed. by S. McWall. New York: St. Martin's Press, 1979.

Houthakker, H. "Compensated Changes in Quantities and Qualities Consumed." *The Review of Economic Studies* 19 (1952): 155–64.

Hsu, F., Anantharaman, T., Campbell, M., & Nowatzyk, A. "A Grandmaster Chess Machine." *Scientific American* 263 (1990): 44–50.

Iacocca, L. *Iacocca: An Autobiography*. Toronto: Bantam, 1984.

Istvan, R. *Business Chess: Perspectives*. Boston: The Boston Consulting Group, 1984.

Itami, H. (with T. Roehl). *Mobilizing Invisible Assets*. Cambridge: Harvard University Press, 1987.

Iwai, K. "Schumpeterian Dynamics." *Journal of Economic Behavior and Organization* 5 (1984): 159–90, 287–320.

Johnson, A. *An Inquiry into the Nature of Value and of Capital and into the Operation of Government Loans, Banking Institutions, and Private Credit*. New York, 1813.

Johnson, G. "Life as a Program on a Cosmic Machine." New York *Times*, February 19, 1989.

Kagano, Tadao, et al. "Comparison of Management Practices between the U.S. and Japan." *Japan Economic Journal*, March, 1983.

Kandinsky, W. *On the Spiritual in Art*. New York: Guggenheim Museum, 1946.

Kant, I. *Collected Works*. Vol. 5. Moscow: Mysl, 1966.

Kantorovich, L. *The Best Use of Economic Resources*. Cambridge: Harvard University Press, 1965.

Kantorovich, L., & Gorstko, A. *Optimal Solutions in Economics*. Moscow: Nauka, 1972.

Katsenelinboigen, A. *Studies in Soviet Economic Planning*. White Plains: M. E. Sharpe, 1978.

———. "L. V. Kantorovich: The Political Dilemma in the Scientific Creativity." *Journal of Post Keynesian Economics* 1 (1978–1979): 129–47.

———. *Soviet Economic Thought and Political Power in the USSR*. New York: Pergamon Press, 1980.

———. *Some New Trends in Systems Theory*. Seaside, Calif.: Intersystems, 1984a.

———. "Planning, Market, and Measurement." In *Acta Slavica Iaponica*. (pp. 1–13). Sapporo, Japan: The Slavic Research Center, Hokkaido University, 1984b.

____. *Vertical and Horizontal Mechanism as a System Phenomenon*. Seaside, Calif.: Intersystems, 1988.

____. *Selected Topics in Indeterministic Systems*. Seaside, Calif.: Intersystems, 1989.

____. *The Soviet Union: Empire, Nation, and System*. New Brunswick: Transactions, 1990.

Katsenelinboigen, A., & Olevskaia, E. "Concerning Some General and Particular Aspects of Different Levels of Management in a Socialistic Economy." In *Modeling Economic Processes*. (pp. 61–101). Moscow: Moscow State University Press, 1972.

King, A. *The U.S. Strategic Minerals Position — The 1980s and Beyond*. Carlisle Barracks: Strategic Studies Institute, U.S. Army War College, 1981.

Knight, F. *Risk, Uncertainty, and Profit*. Chicago: University of Chicago Press, 1971.

Knorr, K. *The Power of Nations*. New York: Basic Books, 1975.

Kolata, G. "Order and Chaos in Computers." *Science* 223 (1984): 917–19.

Kornai, J. *Anti-Equilibrium*. New York: North-Holland, 1971.

Kuznets, S. *Modern Economic Growth*. New Haven: Yale University Press, 1966.

E. Lawber et al., eds. *The Travelling Salesman Problem: A Guided Tour to Combinational Optimization*. New York: Wiley, 1985.

Leibenstein, H. "A Branch of Economics Is Missing: Micro-Micro Theory." *Journal of Economic Literature* 17 (1979): 477–502.

Leontief, W. "A + B = Goodness." *The New York Times Book Review*, January 12, 1986.

Leveson, I., and Wheeler, J. *Economic and Social Changes Influencing U.S. Mobilization Capability*. Discussion paper. New York: Hudson Institute, 1978.

Levine, H. *The Soviet Economy to the Year 2000*. Bureaul of Mines Symposium, May 1981.

Linneman, P. "Some Empirical Results on the Nature of the Hedonic Price Function for the Urban Housing Market." *Journal of Urban Economics* 8 (1980): 47–68.

Little, Arthur D. "Industrial Preparedness in Arms Control Environment." *Summary Report*, ACDH/MEA-246, November 6, 1974.

Luce, R., & Raiffa, H. *Games and Decisions*. New York: Wiley, 1957.

Machlup, F. "Equilibrium and Disequilibrium: Misplaced Concreteness and Disguised Politics." *Economic Journal* 68 (1958): 1–24.

Malenbaum, W. *World Demand for Raw Materials in 1985 & 2000*. New York: McGraw-Hill, 1977.

Mansfield, E., et al. *The Production and Application of New Industrial Technology*. New York: W. W. Norton, 1977.

Materials and the New Dimensions of Conflict: The Implications of Scarce Resources. Carlisle Barracks: Strategic Studies Institute, U.S. Army War College, 1974.

Mayer, C. "Super Fresh Breathes Life into A&P: Storewide Reorganization Plan Spreads into D.C. Area Stores." Washington *Post*, December 8, 1986.

Mayr, E. *The Growth of Biological Thought*. Cambridge: Belknap Press, 1982.

Melman, S. *The Permanent War Economy: American Capitalism in Decline*. New York: Simon and Schuster, 1974.

Mining Annual Review. 1981.

Mintzberg, H. "Planning on Left Side and Managing on the Right." *Harvard Business Review*, July–August 1976, pp. 49–58.

___ *Power in and around Organizations*. Englewood Cliffs, N.J.: Prentice Hall, 1983.

Modigliani, F. "New Developments in the Oligopoly Front." *Journal of Political Economy*, June 1958.

Moffat, D. *Economics Dictionary*. New York: Elsevier, 1976.

Montesquieu, C. *The Persian Letters*. New York: Bobbs-Merrill, 1964.

Myers, K. *Vulnerability of the U.S. to Non-Fuel Mineral Supply Problems*. Philadelphia: University of Pennsylvania, 1983.

National Productivity Review, Summer 1984, pp. 239–45.

Nelson, R., & Winter, S. *An Evolutionary Theory of Economic Change*. Cambridge: Belknap Press, 1982.

Neumann, J., von. "A Model of General Economic Equilibrium." *Review of Economic Studies* 13 (1945): 1–9.

Papp, D. *Soviet Nonfuel Mineral Resources: Surplus or Scarcity?* Bureau of Mines Symposium, May 1981.

Pervozvansky, A. *The Quest*. Moscow: Nauka, 1970.

Petrakov, N., & Rotar, V. *Factor of Uncertainty and Management of Economic System*. Moscow: Nauka, 1985.

Pikler, A. "Utility Theories in Field Physics and Mathematical Economics." *British Journal for the Philosophy of Science*, Vol. 5, 1955.

Polterovich, V. "Concerning Some Abstract Models of Performance of Hierarchically Organized Units in a Controlled Economic System." *Economics and Mathematical Methods* 4 (1968): 176–90.

Pontriagin, L. *The Mathematical Theory of Optimal Processes*. Oxford: Pergamon Press, 1964.

Popper, K. *The Open Universe: An Argument for Indeterminism*. Totowa: Rowman and Littlefield, 1982.

Porter, M. *Competitive Advantage*. New York: Free Press, 1985.

Pryor, F. "A Survey of the Economic Systems of Wild Chimpanzees and Baboons." *Journal of Economic Issues* 15 (March 1981): 33–59.

Ravenschraft, D., & Scherer, F. *Is R&D Profitable?* Cambridge: Strategic Planning Institute, 1982.

The Resource War and the U.S. Business Community: The Case for a Council on Economics and National Security Washington, D.C.: CENS, 1981.

Rizzo, M., ed. *Time, Uncertainty, and Disequilibrium*. Lexington: Lexington Books, 1979.

Rogers, E. *Physics for the Inquiring Mind*. Princeton: Princeton University Press, 1966.

Rosenhead, J. "Planning under Uncertainty." *Journal of the Operational Research Society* 31 (1980): 209–16, 331–41.

Rosenhead, J., Elton, M., & Gupta, S. "Robustness and Optimality as Criteria for Strategic Decisions." *Operational Research Quarterly* 23 (1972): 413–31.

Rozonoer, L. "A Generalized Thermodynamic Approach to Resource Exchange and Allocation." *Automation and Remote Control* 5 (1973): 781–95; 6 (1973): 915–27; 8 (1973): 1272–89.

Samuelson, P. "Economic Frontiers." In *The Collected Scientific Papers of Paul A. Samuelson*. Ed. by J. Stiglitz. Cambridge: MIT Press, 1966, Vol. 2, pp. 1478–92.

____. "Market Mechanisms and Maximization." In *The Collected Scientific Papers of Paul Samuelson*. Ed. by J. Stiglitz. Cambridge: MIT Press, 1949.

Schoeffler, S. *The PIMS Program*. Cambridge: Strategic Planning Institute, 1980a.

____. *Market Position: Build, Hold, or Harvest?* Cambridge: Strategic Planning Institute, 1980b.

____. *Nine Basic Findings on Business Strategy*. Cambridge: Strategic Planning Institute, 1980c.

____. *The Unprofitability of "Modern" Technology and What to Do about It.* Cambridge: Strategic Planning Institute, 1980d.

Schumpeter, J. *The Theory of Economic Development.* Cambridge: Harvard University Press, 1934.

Scitovsky, T. "A Note of Profit Maximization and Its Implications." *Review of Economic Studies* 11 (1943): 57–60.

Seton, F. *Cost, Use, & Value: The Evaluation of Performance, Structure, and Prices across Time, Space, and Economic Systems.* Oxford: Clarendon Press, 1985.

Sevrin, W. *Soviet Nonfuel Minerals: Recent Trends and Prospects.* Bureau of Mines Symposium, May 1981.

Shannon, C. "Programming a Computer for Playing Chess." *The Philosophical Magazine* 41 (1950): 256–75.

Simon, H. *Models of Man.* New York: Wiley, 1957.

____. "Theories of Bounded Rationality." In *Decisions and Organization.* New York: North-Holland, 1972.

____. *Models of Bounded Rationality: Economic Analysis and Public Policy.* Cambridge: MIT Press, Vols. 1 & 2, 1982.

Slutsky, E. "Sulla Theoria del Bilancio del Consumatore." *Giornale Degli Economisti e Rivista di Statistica* 51 (1915): 1–26.

Smets, G. *Aesthetic Judgment and Arousal (An Experimental Contribution to Psycho-Aesthetics).* Leuven, Belgium: Leuven University Press, 1973.

Smilga, V. *In the Pursuit of Beauty.* Moscow: Molodaya Gvardia, 1968.

Strauss, S. *Mineral Self-Sufficiency — The Contrast between the Soviet Union and the United States.* AMC Mining Convention brochure, 1979.

Sukhotin, I., Dementiev, V., & Petrov, A. "On the Category of Efficiency of Production." *Economics and Mathematical Methods* 22 (1986): 126–36.

Szuprowicz, B. *How to Avoid Material Shortages.* New York: Wiley, 1981.

Taylor, H. M. "On the Relative Value of the Pieces in Chess." *Philosophical Magazine* 5 (March 1876): 221–29.

Tsetlin, M. *Research in Collective Behavior of Automata and Modelling Biological Processes.* Moscow: Nauka, 1969.

Tugan-Baranovskii, I. "Influence of Ideas of Political Economy on the Natural Sciences and Philosophy." *Annals of the Ukrainian Academy of Arts and Sciences in the United Sates* 13 (35–36) (1973–1977).

Udis, B. "Economic Conversion." *Science*, May 22, 1987.

Varshavsky, V., & Pospelov, D. *Orchestra with No Conductor.* Moscow: Nauka, 1984.

Volkenstein, V. *Experience of Contemporary Aesthetics.* Moscow-Leningrad: Academia, 1931.

Volkonsky, V. A. "Using the Special Structure of a Matrix to Solve Planning Problems." *Economics and Mathematical Methods* 2 (1966): 235–48.

Wagner, H. *Principles of Operations Research.* Englewood Cliffs, N.J.: Prentice Hall, 1975.

Webster's New World Dictionary. 1980.

Williamson, O. "A Model of Rational Managerial Behavior." In *A Behavioral Theory of the Firm.* Ed. by R. Cyert & J. March. (pp. 237–52). Englewood Cliffs, N.J.: Prentice Hall, 1963.

Young, W. "Time, Measurement, and Indeterminacy in Economics." In *Quality and Quantity.* (pp. 455–62). Amsterdam: Elsevier, 1982.

Zeleny, M. *Multiple Criteria Decision Making.* New York: McGraw-Hill, 1982.

Zelikina, L. "Macroeconomic Models of Optimal Investments in Technological Progress and Turnpikes Theorems." *Economics and Mathematical Methods* 11 (1975): 453–67.

Zhabin, A. "Is There an Evolution of Minerals on Earth?" *Science and Life*, No. 2, 1980.

Author Index

Ackoff, Russell, 3, 12, 33, 35, 65–67, 79, 83, 86, 110, 161, 162, 235, 262–65, 268–70, 279, 285
Alexander II, 203
Ambartsumian, Victor, 214
Antonov, Oleg, 214
Arrow, Kenneth, 84
Ashby, W. Ross, 28
Asoka, 88

Bakhtin, Michail, 302
Barone, Enrico, 115
Bartol, Julio R., 270, 299
Baumol, William, 3, 15, 250, 271–73, 280, 281
Becker, Abraham, 180, 182
Bergson, Abram, 169, 182
Bergson, Henri, 26
Birkhoff, George, 157–60
Borch, Fred, 254
Bosch, Hieronymus, 36
Botvinnik, Michail, 15
Boulding, Kenneth, 1, 21, 22, 35, 163, 170
Bowman, Edward, 284
Boyai, Farkos, 7
Brahe, Tycho, 114
Buddha, 88
Buzzell, Robert, 3, 257

Cavendish, Henry, 212
Cezanne, Paul, 212
Chukhrai, Grigory, 247, 248
Cyert, Richard, 3, 250, 282, 283

Danilov-Danil'yan, Victor, 121
Dantzig, George, 8, 115, 116, 128, 129, 247
Dostoyevsky, Fedor, 302
Drucker, Peter, 3, 15, 250–52, 262, 263, 271, 274–79
Dudkin, Lev, 127, 172
Dynkin, Eugene, 78, 130

Einstein, Albert, 7, 25
Elton, Martin, 132
Emery, Fred, 86, 236
Ershov, Emil, 127
Euler, Leonard, 114
Farjoun, Emmanuel, 2, 22
Fedorov, Nikolai, 219
Feynman, Richard, 106
Fine, Daniel, 226
Friedman, Joel, 163

Galbraith, John Kenneth, 280
Gale, Bradley, 3, 255
Galois, Evariste, 84
Gardiner, Ralf, 251
Gauss, Carl Friedrich, 7
Godel, Kurt, 84
Gelfand, Israel, 78
Georgescu-Roegen, Nicholas, 2, 56
Geranmayeh, Ali, 270, 299
Gerschenkron, Alexander, 180
Gharajedaghi, Jamshid, 12, 23, 35, 57, 58
Gorbachev, Michail, 203, 208, 247
Gordon, Robert, 250
Griliches, Zvi, 151, 177
Gross, Norman, 288
Gupta, Shiv, 132

Hegel, Georg, 61
Heinman, Solomon, 13
Heinrich, B., 20
Houthakker, Hendrik, 151

Iacocca, Lee, 236, 274
Istvan, Rudyard, 15, 252
Itami, Hiroyuki, 3, 243
Iugenburg, Semen, 14
Ivan the Terrible, 203
Ives, Charles, 213

Johnson, Alexander, 213

Kandinsky, Wassily, 36, 185
Kant, Immanuel, 61
Kantorovich, Leonid, 8, 115, 116, 123, 124, 130, 131, 170, 246
Kepler, Johan, 114
Khrushchev, Nikita, 86
King, Alvin, 3, 196, 225
Kolmogorov, Andrey, 48, 70, 78
Koopmans, Tjalling, 8, 115, 116
Koriakin, Yury, 107
Kornai, Janos, 3, 247
Korolev, Sergey, 219
Kravis, Irving, 180
Kuznets, Simon, 3, 182–91

Lagrange, Joseph Louis,114
Lange, Oscar, 215
Lebedev, Sergey, 214
Leibenstein, Harvey, 3, 234, 247
Lenin, Vladimir, 85, 86
Leontief, Wassily, 1
Lerner, Abba, 215
Leveson, Irving, 3, 196
Levine, Herbert, 222
Linneman, Peter, 152
Lisowski, John, 163
Lobachevsky, Nikolai, 7
Long, Alton, 296

Machover, Moshe, 2, 22
Malenbaum, Wilfred, 221, 222
Mao Tsetung, 86
March, James, 3, 250, 282, 283
Marx, Karl, 41, 85, 116, 273
Maupertuis, Pierre, 114
Mayr, Ernst, 120
Melman, Seymour, 3, 203, 206–8
Minkovsky, Herman, 115
Montesquieu, Charles, 296
Morgenstern, Oscar, 115
Muraviev, Vladimir, 214
Myers, Kent, 163, 221, 226

Nelson, Richard, 2, 4
Neumann, John, 8, 97, 115, 173, 179
Newton, Isaac, 7, 38, 114
Nicholas II, 203
Nixon, Richard, 218

Papp, Daniel, 222, 223
Pascal, Blaise, 48
Pervozvansky, Anatoly, 78
Peter the Great, 203
Petrakov, Nikolai, 131
Polterovich, Victor, 14, 235
Popper, Karl, 38, 83
Porter, Michael, 3, 15, 273
Pryor, Fred, 20

Radec, Karl, 86
Ramsey, Frank, 115
Riasanovsky, Alexander, 213
Riemann, Bernhard, 7
Rineherd, James, 35
Rosenhead, Jonathan, 132, 134
Rotar, Vladimir, 131
Rozonoer, Lev, 8, 20, 78

Samuelson, Paul, 121, 168
Schultz, Jacob, 163
Schumpeter, Joseph, 59, 78, 278
Scitovsky, Tibor, 236
Seton, Francis, 4, 180, 181
Severin, William, 223, 224
Shemiakin, Michail, 36
Schoeffler, Sidney, 3, 254–59
Simon, Herbert, 3, 15, 45, 52, 55
Slutsky, Eugene, 102, 110, 115
Smilga, V., 7
Stalin, Joseph, 13, 86, 121, 190, 203, 246
Stevens, Wallace, 213
Strauss, Simon, 223, 224
Suslov, Michail, 248

Taga, Toshiaki, 299
Taub, Moses, 13
Taylor, Fred, 216
Tsetlin, Michail, 127
Tsiolkovsky, Konstantin, 219
Tugan-Baranovskii, Michail, 27

Van Gogh, Vincent, 35
Volkenstein, Vladimir, 73
Volkonsky, Victor, 127

Wheeler, James, 3, 196
Wilde, Oscar, 13
Williamson, Oliver, 236
Winter, Sidney, 2, 4
Walras, Leon, 111, 115

Young, W., 2

Zavel'sky, Michail, 121
Zelikina, Ludmila, 97, 177

Subject Index

academy, 32, 122, 124, 213
abstract paintings, 36, 165
adaptation, 51, 131
aesthetics, 59–63, 73, 74, 146, 153, 158–63, 214, 295, 298; and art, 13, 36, 157, 160, 212, 213; beauty, 61, 62, 73, 74, 157, 160–63, 296; harmony, 36, 61, 158; measurement, 158–62
analysis, 1, 12, 62

biological systems, 15, 31, 120, 123
Buddhism, 87, 88, 210

cancer, 218, 219
chess, 15, 33, 47, 55, 69–71, 73, 75, 77, 79, 154, 160, 164, 244, 252, 253; combinational style, 73, 279; computer programs, 15, 55, 70, 153; positional style, 14, 69, 71, 73, 153, 279; retrograde chess, 85
China, 85, 86, 90, 116, 134, 194
church, 122, 294
competition, 31, 32, 210, 211, 216, 222, 240, 253, 254, 256, 257, 265, 272–74, 277, 279, 284, 288
complexity, 57, 58, 158, 159, 284
computers, 11–13, 44, 47, 89, 115, 116, 131, 170, 294, 300. *See also* chess (computer programs)
creativity, 12, 13, 25–27, 30, 43, 59, 72, 73, 88, 146, 151, 154, 213, 214, 216, 220, 274, 294
criteria of optimality: objective function, 103, 107, 138, 193; profit, 3–6, 10, 11, 15, 66, 75, 76, 118, 128, 129, 132, 138, 156, 159, 162, 165, 169, 205, 206, 234–41, 245–48, 251, 252, 254, 255, 257, 260, 261, 264, 265, 271–73, 277, 278, 380–83, 297; profitability, 116, 246, 252–60, 264; utility function, 57, 78, 97, 126, 139, 166, 167
cultures: extrovert, 27, 57, 60, 87–91, 190; introvert, 27, 60, 87–91, 190

Darwinism, 31
deduction, 30, 93, 112–14, 148, 164, 177, 258, 291, 292
definition, 33–35, 42; alternatives (dilemmas), 63, 195, 210, 240, 296; corporate, 10, 35, 156, 157, 159, 238, 241, 242, 261, 283, 285, 299; criteria, 56–58, 65, 97, 104, 105, 109, 117, 202, 241; driving forces, 97; measurement, 15, 30, 73, 148, 173–83, 191, 240, 251, 298, 299; path (course), 62, 63, 76, 117–19, 138, 147, 260, 275, 287–89, 292, 297, 298; scientific-technological, 5, 32, 88–89, 112, 171, 174, 176, 204, 208, 210, 211, 219, 259, 274, 292, 297; starting points, 29, 30; two-fold sources, 28–31, 36; unexpected (unforeseen) outcomes, 40, 55
development, 3, 4, 6, 25, 28, 29, 33–35, 57, 60, 67, 85, 89, 95, 116, 117, 120, 121, 123, 128, 166–72, 187, 189, 191, 194, 199, 202, 212, 215, 241, 259, 270, 277, 285, 292;
dynamics of economic systems, 25, 27, 33, 58, 59, 97, 117, 118, 120, 146, 169, 173–75, 177, 179, 181, 183, 187, 239, 298; "modern economic growth," 3, 182–90; survival, 3, 21, 31, 33–35, 124, 134, 166–68, 170, 172, 218, 252, 267, 283. *See also* development, growth

economic system: definition, 19–23; normal-pathological, 23, 24
emotion, 34, 107, 124, 159, 162, 163, 165
ends: eschatology, 124; telelogy, 117, 119, 120; teleology, 110, 117–20; teleomatic, 120; teleonomic, 120; telology, 117, 120

entropy, 56–59, 78
equilibrium, 2, 52, 58, 59, 75, 98–102, 106, 107, 109–11, 126, 130, 151, 166, 168, 170, 173, 238, 249, 282; definition, 96, 97, 106; disequilibrium, 4, 58, 59; dynamic, 120, 173–75, 177; models, 6, 15, 78, 96, 99, 101, 104, 105, 110, 111–17, 119, 120–22, 150, 166, 181, 222, 302; saturation, 98
Euclid's fifth postulate, 7
evolution, 25–27, 34, 36, 37, 58, 89, 98, 109, 119, 165, 292

God, 55, 119, 122, 182, 296; demiurge, 43, 99, 117
Great Leap Forward, 86
growth, 3, 24, 33–35, 66, 86, 96–99, 166–69, 183; consumer goods, 108, 201; industrial production, 90, 205, 247; measurement, 4–10, 14, 15, 56, 169–82, 184, 185, 239, 297–99; military power, 194, 203, 206; population, 120, 185–87; sales, 251, 256, 257, 271, 272; standard of living, 150, 195, 203;

Harvard School of Business, 255
heuristic, 25, 37, 61, 66, 123, 138, 145, 248–53, 262, 263
hippie movement, 88

impossibilities, 44, 46, 47, 60, 83, 84, 147, 156; perpetual motion machine, 83
indeterminism: degree, 5, 11, 12, 56, 59, 61, 68, 111, 297; foundations of economics, 1–3, 5, 12, 15, 16, 111, 138–40, 146–48, 177, 179, 181, 194–96, 212, 216–19, 245, 247, 248, 261, 281, 284, 297, 302; philosophy, 12, 36–56, 110, 139, 163, 164, 177
induction, 93, 112–14, 126, 177, 258
Institute for Interactive Management (Interact), 299
Integration: horizontal, 281; vertical, 156, 242, 254, 256, 280

Japan, 59, 90, 169, 176, 179, 209–11, 243

Lagrange function, 100, 112, 114
Lagrange multipliers, 114, 148, 175
large scale problems, 54, 75, 76, 91, 93, 115, 125–27, 130, 137, 150, 172, 215, 216, 238, 239, 247, 272, 284; abregation, 127, 132–34; abstraction, 127, 132, 133, 150; aggregation-disaggregation, 55, 66, 71, 109, 110, 126, 127, 132, 133, 150, 165, 169, 170, 172, 176, 177, 179, 222, 253, 258, 271, 282. *See also* mathematics (decomposition)
laws of: conservation of energy, 96–98, 106; conservation of matter, 95–98; diminishing return, 97; economics, 115–17, 128, 198, 293; excluded middle, 68; nature, 38, 40, 43, 44, 61, 106, 113–16, 119, 124, 126, 164, 170, 257–59, 286
leaders: administrator, 291, 293–95; entrepreneur, 59, 78, 278, 280, 283, 290–92; manager, 161, 162, 165, 205, 206, 236, 239, 243, 245, 247, 252–55, 259, 261, 264, 267–69, 282, 283, 290–92, 294, 295, 299–302; "navigator," 291, 292, 296

market, 2, 21, 22, 29, 31, 65, 78, 104, 109, 117, 119, 121, 122, 133, 145, 152, 176, 182, 198, 203, 205, 212, 214, 217, 222, 224, 246, 247, 251, 254–57, 265, 267, 272, 275, 276, 284, 295, 297, 300; bazaar, 78; laws, 257–59; market share, 10, 146, 156, 242, 243, 250, 251, 256, 257, 259, 260, 265, 282, 300; marketing, 76, 165, 205, 222, 243, 254, 267, 276, 277, 283; new products, 109, 150, 211, 252, 256; nonmarket, 78, 215; stock market, 5, 65, 139
Marxism, 24, 84, 85, 87, 119, 122, 215
mathematical economics, 10, 78, 122, 124, 163, 216, 234, 237
mathematics, 7, 8, 19, 31, 32, 36, 48, 75, 84, 115, 122, 160, 184, 213, 219, 262, 278, 302; decomposition, 54, 125–30, 216, 247; dynamic programming, 66, 128; "knapsack" problem, 130; linear programming, 8, 54, 126, 131, 170; programming, 8, 112, 170, 216
mechanisms of performance: horizontal, 22, 29, 59, 78, 104, 118, 126, 127, 131, 212, 215, 217, 281; vertical, 22, 29, 59, 65, 104, 118, 126, 127, 130, 212, 215–27. *See also* market, planning
methods of performance: belief, 40, 59–61, 66, 84, 86, 138, 190, 226, 228, 232, 233, 294, 297; directing, 12, 62–64, 67, 68, 83, 138, 274, 290, 292, 295, 297; faith, 27, 68, 84, 86, 105, 109, 123, 294, 295; programming, 12, 49, 51, 54, 62–64, 67, 68, 72, 74, 76, 77, 91, 108, 124, 125, 129, 132, 148, 149, 153, 168, 196, 206, 207, 213, 215–18, 221, 222, 225, 245, 292, 293, 295, 296, 299, 300, 302; scientific, 25, 44, 59–63, 77, 83, 119, 122, 162, 192. *See also* aesthetics, potentiating method

modes of performance: reactive, 4, 6, 43, 44, 61, 93, 110, 121, 125, 126, 130, 138, 145, 172, 221, 248, 263, 298; selective, 43, 44, 61, 94, 110, 124, 126, 130, 138, 145, 172, 173, 221, 298, 299

national income, 4, 169, 171, 172, 179–81, 184, 191, 222, 238

objectivity, 48, 60, 61, 68, 123, 244, 261, 292–94
order, 5, 51, 56–59, 61, 93, 109, 158, 159, 216, 270, 272, 284, 297
outerspace program, 217–219
ownership, 41, 212, 261; foundations, 214, 218; stockholders, 280, 282, 288–90; venture capital, 266, 295

parameters: conjugated, 2, 7, 8, 92, 93, 97, 98, 100, 104, 106, 107, 112, 139, 146–49, 164, 181, 242, 275; essentials, 69, 75, 76, 91, 139, 146–50, 157, 158, 179, 255, 258, 282; material, 15, 35, 69–72, 74, 76, 91, 139, 145–47, 149, 153, 154, 157, 159, 163, 167, 168, 178, 179, 181, 183–87, 191, 197, 198, 200, 203, 204, 238, 242, 244, 248, 253, 259–61, 263–65, 269–71, 274, 275, 281, 282, 290, 300; primary, 2, 71, 66, 77, 98, 106, 107, 112, 120, 139, 146, 147, 163, 225, 254, 275, 276; rank, 107; relations, 69, 70, 139, 145–50, 152, 156, 157, 159, 163, 164, 167, 168, 181, 184–86, 191, 198, 200, 222, 225, 238, 241–45, 248, 253–55, 258–65, 268, 269, 273, 275, 276, 278, 280–82, 299
participation, 162, 270, 285, 288, 290
PERT, 270
physical systems, 58, 106, 114, 170; extremal principle, 122, 123; Plank's constant, 164; quantum mechanics, 2, 42, 164; relativity theory, 7, 25; statistical mechanics, 2; thermodynamics, 78, 83; uncertainty principle, 164
"piecemeal engineering," 83
PIMS, 146, 243, 253–61, 300
pioneer, 213, 214, 219, 278, 294, 295
pioneering, 3, 5, 14, 30, 151, 176, 210, 211, 214, 217, 250, 261, 271, 278, 292, 294, 296
plan, 58, 59, 66, 71, 85, 86, 128, 131–33, 182, 192, 193, 200, 202, 207, 246, 270, 287–89, 292, 293, 299; centrally planned systems, 21, 22, 78, 117, 119, 121–23, 133, 182, 214, 215, 224, 245, 246
planning, 29, 64–68, 79, 104, 117, 120, 122, 137, 139, 165, 200, 246, 255, 279, 299; contingency type, 50, 131; horizon, 28, 239–41, 243–45, 255, 260, 261, 288; integrative type, 65; interactive type, 65; long term time, 30, 34, 68, 133, 146, 159, 185, 189, 194, 203, 251, 254, 260, 288–90; long-long range time, 68; mid-term time, 67, 68, 222, 245; operational time, 67, 68, 264; optimal type, 13, 14, 100, 123, 246; perspective time, 67, 68; planners, 66, 246, 254, 260, 299; prospective time, 67, 68; schedule type, 49; short term time, 133, 159, 220, 230, 251, 254, 272; strategic time, 67, 68, 71, 132, 243, 254–57, 265, 269, 270, 288, 289, 299, 300; tactical time, 67, 68, 132, 299; team planning, 288
political system features: democracy, 84, 118, 203, 206, 284–86, 295; ochlocracy (mobocracy), 203, 285, 286, 295; openness, 203, 284, 286, 287, 295; pluralism, 19, 63, 64, 203, 206, 219, 284, 286, 287, 289, 290, 293, 295; separation of power, 122, 203, 284, 286, 289, 295
political systems, 284, 286, 295, 296
political system types: capitalism, 24, 116, 128, 129, 208; communism, 26, 83–86, 116, 118, 188, 189; fascism, 83; feudal, 210, 283–85, 292; socialism, 24, 85, 122, 124, 128, 194, 215, 245
potential, 30, 33, 52, 53, 61, 62, 64, 67, 133, 137–39, 145, 147, 174, 184, 190, 191, 274, 283, 299; corporate, 248, 251, 252, 256, 261, 295, 296; definition, 166–70, 181, 182, 256; measurement, 51, 62, 145, 164, 166, 170, 178, 182, 183, 185; military, 192–94, 196–207
potentiating method, 12, 48, 62–64, 67–69, 72, 76, 137–39, 146, 148, 149, 153, 156, 163, 164, 234, 261, 270, 279–83, 292,

295, 298, 299, 300, 302; pathological, 245; principle of asymmetry, 271, 273; quasi, 281
presentation of an economic system: analytical, 6, 60, 62, 93, 100–2, 110, 112–16, 119, 120, 122, 126, 172, 297; differential, 95, 96, 120, 238, 248, 271; integral, 95, 98, 99, 111, 120, 238, 239, 241, 242, 248; optimality, 6, 15, 98–105, 109–12, 114–22, 128, 132, 150, 166, 171, 173–75, 177, 181, 182, 272, 297; tabular function, 93, 94, 101, 109, 110, 112–14. *See also* equilibrium, Lagrange function
prices: complete-incomplete, 147, 237, 238, 248; constant, 6, 169, 175, 180, 297; current, 4–6, 75, 180, 184, 260, 261, 271, 297; equilibrium, 153, 175, 176, 182; hedonic, 150–53; unconditional, 147, 152, 154, 156, 157, 167, 178, 179, 242–44, 261, 271
private goods, 210, 211
production function, 60, 93, 94, 115, 177
psychology, 32, 35, 56, 57, 88, 95, 107, 115, 126, 145, 160, 165
public goods, 176, 210, 211
purpose, 3, 52, 55, 65, 79, 90, 121, 173, 188, 201, 202, 255; aim, 11, 35, 52, 66–68, 71, 77, 79, 138, 183, 223, 224, 260, 271, 273, 274, 283, 295; goal, 29, 61, 64–68, 71, 72, 74–77, 79, 96, 106, 117–20, 128, 151, 153, 154, 162, 166, 178, 192, 195, 199, 200, 202, 203, 215, 217–19, 251, 252, 255, 264, 268–71, 274–77, 280–83, 295, 298; ideal, 58, 66–68, 79, 83–86, 88, 117, 192, 215, 243, 266, 269, 270, 287; objective, 10, 19, 34, 52, 65–70, 72, 74, 76, 79, 118, 128, 133, 148, 164, 193, 215, 243, 245, 251, 254, 257, 269, 270, 276, 282, 283, 299; target, 66, 68, 72, 76, 79, 271, 281
Pyrrhic victory, 66, 200, 261

research and development (R&D), 23, 30, 126, 176, 177, 209, 211, 228, 255, 256, 259
reserves, 13, 43, 49, 51, 131, 134, 196, 198, 199, 222, 227, 242
resilience, 51
Ricardian principle, 210
robustness, 51, 132, 134

satisficing, 52
social systems, 20, 23, 24, 149
South Korea, 211
Soviet Union (USSR), 13, 24, 25, 90, 121, 122, 128, 180, 189, 192, 194, 195, 200–3, 205, 208, 214, 215, 221, 233, 245, 246, 296
statistics, 14, 93, 109, 145, 151, 177
strategic minerals, 3, 139, 146, 220–33
Strategic Planning Institute (SPI), 255
subjectivity, 48, 60, 70, 138, 244, 252, 300
supply and demand, 100–3, 106, 107, 110, 115, 166, 220; demand, 2, 6, 54, 95, 102, 103, 105, 107, 110, 133, 187, 189, 200–2, 204, 205, 215, 220, 222–24, 265, 274, 283, 297; supply, 89, 96, 133, 139, 187, 188, 220, 222, 225, 227, 230, 252, 268
synthesis, 36, 62, 63, 65, 103–8, 127, 138, 151, 242, 244, 270
systems approach, 11, 12, 19, 20, 35, 37, 38, 42, 47, 48, 96–98, 138, 167, 191, 236, 292
systems structures: closed, 43, 44, 46, 56, 78, 117, 118, 178; hierarchy, 44–46, 53, 129, 137, 163, 168, 200, 215, 216, 284; "lowerarchy," 285; open, 43, 56, 91, 173, 175, 179, 203; purposeful, 65; self-acting, 7, 20; subsystems, 53–56, 124, 127–29, 138, 139; suprasystems, 12, 19, 20, 27, 53–55, 57, 105, 107, 113, 115, 236, 292

Taiwan, 211
tenure, 213
theology, 15, 63, 119, 124
turnpike theorem, 120, 146, 177, 179, 243, 253

uncertainty: 40, 42, 44, 48–51, 52, 56, 69, 106, 109, 129–32, 134; certainty, 2, 40–42, 44, 48, 49, 51, 52, 64, 66, 69, 125, 129–31, 241, 298; chaos, 5, 52, 53, 297; complete uncertainty, 40, 51, 52; degree, 38, 40, 42, 48, 49, 53, 129; predispositions, 52, 53, 61, 62, 67, 71, 73, 138, 139, 147, 148, 150, 156, 159, 160, 164, 165, 196, 197, 225, 240, 241, 258, 298–300, 302; probability, 2, 6, 10, 48–53, 60, 62–64, 71, 78, 129, 130, 131, 134, 148, 164, 238, 241, 297, 298, 302
unconditional values, 15, 69–72, 76, 147, 149, 150, 154, 166, 167, 181
uniqueness, 51, 62, 103, 148, 181, 244, 245, 298

universities, 32, 78, 122, 176, 213, 219, 220, 266, 268; Carnegie-Melon, 282; University of Pennsylvania, 35, 165, 180, 213, 269
utopia, 25, 83, 88, 117, 203

vulnerability, 3, 51, 193, 198, 225–33

About the Author

ARON KATSENELINBOIGEN was born in the Soviet Union and studied at the Moscow State Economic Institute. Since his emigration to the United States, Dr. Katsenelinboigen has taught at Princeton University, and since 1978, the University of Pennsylvania's Wharton School. He is the author of 14 books including, most recently, *Selected Topics in Indeterministic Systems* (1989) and *The Soviet Union: Empire, Nation, and System* (1990).

1-1 8

WHARTON REPROGRAPHICS

Paul R. Kleindorfer Howard C. Kunreuther
Paul J. H. Schoemaker

Decision Sciences
An integrative perspective

1
The scope of decision sciences

CAMBRIDGE
UNIVERSITY PRESS

1.1 Introduction

The emerging field of decision sciences is concerned with understanding and improving decision making of individuals, groups, and organizations. At issue is not only how decision makers "solve problems," but also how they came to identify and accept such problems and learn from the results of their actions.

Decision making may be defined as intentional and reflective choice in response to perceived needs. Anthropologists such as Pierre Teilhard de Chardin (1959) consider this ability to reflect and choose as the fundamental characteristic distinguishing man from lower forms of life. While the recognition of this ability may provide the reader with a momentary feeling of superiority, history is also replete with examples of human limitations in decision making. Our purpose in this book is to provide a basis for understanding these limitations as they relate to individual, organizational, and societal decision processes and to use this basis to provide insights as to how such decision processes might be improved.

The dilemmas of choice in the face of an uncertain and complex world have long been the focus of religion, literature, and philosophy. In Western literature, classical epic poems and tragedies depict the gradual evolution of human choice from a metaphorical extension of the will of gods in Homer's epics to the realm of willful, if not always rational, choice in Euripides' tragedies. Writers such as Friedrich Nietzsche and Julian Jaynes have described this emergence during the first millennium B.C. from ritual and metaphor to willful action and reflective choice as a matter of the utmost importance in Western culture. People began to view their actions as something beyond the momentary impulses of their current existence to an evolving, connected whole of past, present, and future. The result of this was the view that people could partially affect, and be held responsible for, their choices. It was to be many centuries, however, before the process of reaching such choices was subjected to systematic analysis.

Religion and moral philosophy provided the early foundations for analyzing choice behavior by setting rules for evaluating the motives and

	DESCRIPTIVE THEORIES	PRESCRIPTIVE THEORIES
Individual	Psychology Marketing Psychiatry Literature	Decision Theory Economics Operations Research Philosophy/Logic
Group	Social Psychology Organizational Behavior Anthropology Sociology	Game Theory Organization Behavior Clinical Psych/Therapy Finance/Economics
Organization	Organization Theory Sociology Industrial Organization Political Science	Planning – Strategy Control Theory/Cybernetics Organization Design Team Theory/Economics
Society	Sociology Anthropology Macro Economics	Legal Philosophy Political Science Social Choice

Figure 1.1. The disciplinary roots of decision sciences

consequences of choice. But it was not until the triumphs of natural science in the seventeenth century that a basis was laid for the systematic investigation of natural phenomena and, much later, of human behavior.

Over the next two centuries, fundamental advances were made in philosophy, economics, biology, psychology, and sociology. Taken together, the result of these studies was a model of the human species as a complicated, evolving organism, embedded in a series of nested social and economic systems. Decision making and choice became areas of immense theoretical and applied interest in many fields. The elaboration and integration of these fields continue. Figure 1.1 lists some of the current disciplines involved in the study of decision sciences today. At the risk of oversimplification, we have divided these fields into two categories of analysis: descriptive and prescriptive. Descriptive analysis refers to how people actually make decisions. Prescriptive analysis indicates how decisions should be made according to a set of well-defined criteria. Prescriptive analysis can range from formal axiomatic theories to more informal approaches for aiding decision making. As we shall see, which prescriptive approach one selects for improving the decision process requires understanding the context and descriptive realities of the situation in question.

At the descriptive level, research in the behavioral sciences has increased our understanding of the role of organization and structure on the decision making process. In particular, significant empirical efforts have been undertaken on procedures used by organizations to resolve conflicts and generate solutions between decision makers with different goals and objectives. Political science has addressed similar issues at the level of governmental and bureaucratic decision making. Psychology provides experimental evidence on the limited ability of decision makers to process information and the systematic biases and simplifying decision procedures people employ. Finally, sociology provides a perspective on the nature of information diffusion processes and the role of social norms in the decision-making process.

At the prescriptive level, economics offers a well-developed paradigm for structuring decision making by individuals and firms. At the societal level, an extensive literature is now emerging that provides guidelines for aiding the public choice process. The fields of management science and operations research arose during World War II and thereafter, with an initial emphasis on improving decision making in military and business organizations. A wide range of optimization models has been developed for the purpose of maximizing profits or minimizing costs for part of the firm's operations. In recent years these techniques have been extended to many areas of societal decision making, such as transportation and energy system planning. Finally, the field of information systems is a product of the computer age. Its primary contribution to decision sciences is in the development of interactive computer models, sometimes referred to as decision support systems, which improve decision makers' ability to evaluate complex decision situations.

In reality, it is not so straightforward to categorize a theory as descriptive or prescriptive. Psychology has been primarily concerned with characterizing individuals' information-processing abilities and identifying systematic biases through field studies and controlled laboratory experiments (descriptive analyses). Based on this understanding of human behavior, there has been research by psychologists in trying to correct these biases (prescriptive analyses). Until recently, research in economics focused almost exclusively on the development and application of models of choice as to how individuals and firms should behave (prescriptive analysis). Today more attention is being given to understanding the impact that imperfect information and misprocessing of data have on the final choices of individuals and the performance of markets (descriptive analysis).

Utilizing concepts from these disciplines, we view the field of decision sciences as integrating descriptive analysis with prescriptive recommendations. Descriptive analysis provides insights on the values, beliefs, and choice processes actually used by decision makers. Prescriptive analysis helps to clarify values and beliefs and suggests ways of improving decision processes using criteria from rational models of choice, such as expected utility theory, as guidelines. The basic framework of this book for integrating descriptive and prescriptive analysis is described at the end of this chapter. We first consider some of the important organizational and technological factors that have affected decision making in industrial societies and then illustrate the problems of interest in this book with some examples.

1.2 Historical developments

It is instructive to consider from a historical perspective why the study of decision making has become increasingly important. The reasons are evident. During the two centuries since the industrial revolution, goods and services available in the market have grown exponentially in variety and complexity. Revolutionary changes in transportation and communication have made economies, both national and international, more tightly interconnected. The workplace has become the realm of science and electronics in place of the manual arts and crafts of two centuries ago. The buildings we live and work in are filled with a bewildering array of laborsaving innovations, which are wonderful when they function properly, and which remind us of an earlier, simpler age when they don't. When one considers all of this, it is clear that the pace of life has quickened. There has been a tremendous increase in the complexity and frequency of decisions confronting individuals and organizations as they seek to understand and cope with their social and economic environment.

Consider individual decision making. Two centuries ago, most decisions revolved around one's survival and associated daily family activities (excluding a few aristocrats who had other concerns). The situation is vastly different now. Today's teenager is already making decisions on complicated consumer items such as cars and computers that were not even invented a century ago. The price of all this is longer educational preparation for life and a much deeper division of jobs and expertise once maturity is reached. Thus, as Alvin Toffler has pointed out, our personal lives have become exceedingly more complex in terms of the choices we must make and, at the same time, our educational and professional lives have become narrower as we attempt to stay abreast of at least one of the myriad occupational specialties responsible for supporting our lives.

There have been both positive and negative consequences of these changes at the individual level. Very few would gainsay the importance of the increases in the quality of life and level of goods available to the average person today relative to his predecessors of a century ago. On the other hand, the dizzying pace of associated choice activity and the specialization in education and jobs associated with today's high technology organizations have also caused a deep feeling of loss of control and alienation among many members of modern society. The post–World War II mobile society and the liberation of women have had their impact on the family and have exacerbated the feeling of being alone in a world that one cannot control. Leaving the nuclear family and living in a society that can be wiped out in a nuclear war has created its own set of tensions.

One interesting development of this felt alienation has been a growth of interest in group choice processes as a means of sharing responsibility for these complicated decisions and partially recapturing the emotional support that the family provided in the past. The number of community, social, and economic groups supporting individual interests has grown astronomically over the past century as individuals have attempted to gain a sense of identity and control over their destinies through banding together with like-minded others. Interestingly, although these groups have provided support to individuals, the sheer number of such groups has immensely complicated the social and political processes attempting to balance individual interests in setting public policy.

The recent controversy in different states on the abortion issue stimulated by the Supreme Court decision is one of many examples of the importance of groups in rallying around certain positions and making a political impact. Mancur Olson (1965) has described the increasingly complex problem of dealing with such special interest groups as a major impediment to national economic growth and social consensus. Thus, some of the steps taken by individuals to cope with increased choice complexity have had far-reaching effects on decision making at the group and societal levels as well.

These same effects are also evident in organizational decision making. Two centuries ago, all but a few governmental organizations were small in size with only rudimentary administrative and accounting procedures. The industrial revolution, starting in England in the late eighteenth century and spreading like wildfire to Europe and the United States, changed all of this. First, factories were built to house the newly developing mass production technologies. New energy sources were developed to power these complexes. Organizational and management innovations were designed to cope with the increasing size and complexity of these factories.

The economies of scale available through these new manufacturing technologies provided incentives to develop wider markets. New transportation modes, both rail and shipping, made this possible, and brought with them at the end of the nineteenth century the beginnings of the large geographically dispersed firms that have grown into today's conglomerates.

Alfred Chandler (1977) has called this transformation from small, family-centered business units to multinational giants the substitution of Adam Smith's "invisible hand" of the marketplace by the "visible hand" of internal organization. Chandler ascribes this substitution to the fact that the complex products we have come to rely on in modern society require much larger production and service organizations than the simpler products of former times. These large organizations also imply immense problems in coordinating and controlling organizational decision making, and in designing the incentive and information systems that support such decision making.

At the societal level there has been a radical change in the role that government has played in our lives in the past fifty years. Prior to that time individuals and firms operated in a relatively laissez-faire environment with the government collecting taxes to support a relatively small bureaucracy and military activities, and with firms occasionally prosecuted for violating antitrust regulations. Today there are a number of federal and state agencies that are involved in information provision and regulation to protect us against health, safety, and environmental risks. Financial risks are mitigated through unemployment insurance and government-subsidized health care. How much risk we are willing to tolerate is a subject of considerable controversy. There is a price to pay for zero risk. The challenge is to frame the problem so people can see what the trade-offs are.

This brief historical sketch points out the vastly changed role of consumers, organizations, and social institutions over the past two centuries. A primary goal of this book is to understand the implications of these changes for decision making, both descriptively and prescriptively.

1.3 Framework of the book

The preceding discussion illustrates the broad scope and importance of decision sciences. In this section we wish to outline briefly the framework this book follows in addressing these issues.

1.3.1 *Key aspects of decision making*

Figure 1.2 depicts two major areas of interest with respect to the decision-making process. The box at the top, labeled procedural and technological

8.2

THEORIES OF BOUNDED RATIONALITY

HERBERT A. SIMON

1. Introduction. – 2. Approaches to rational choice in chess. – 3. Bounded rationality in design. – 4. Bounded rationality in management science. – 5. Conclusion.

1. Introduction

Rationality denotes a style of behavior that is appropriate to the achievement of given goals, within the limits imposed by given conditions and constraints. Theories of rational behavior may be normative or descriptive – that is, they may prescribe how people or organizations should behave in order to achieve certain goals under certain conditions, or they may purport to describe how people or organizations do, in fact, behave. This essay will be concerned with the structure of theories of rational behavior, whether they are intended prescriptively or descriptively.

Individual and organizational rationality. A theory of rational behavior may be concerned with the rationality of individuals or the rationality of organizations. In fact, the two bodies of theory are not wholly distinct.[1] One plausible distinction between them is that a theory of organizational rationality must treat the phenomena of goal conflict, while a theory of individual rationality need not. This is only partly correct, for goal conflict may be important in individual as in group behavior – it is a major theme of so-called "dissonance theory" in psychology. (See N. P. CHAPANIS and J. A. CHAPANIS (1964).) A theory of individual behavior microscopic enough to concern itself with the internal organization (neurological or functional) of the central nervous system will have a significant organizational component. A theory of organizational behavior macroscopic enough to treat the organization as a monolith will be a theory of an "individual." Although this chapter will be aimed primarily at understanding individual rationality, I shall not hesitate to use the theory of the firm – classically, the theory of a monolithic entrepreneur – as a convenient and enlightening illustrative example.

---

[1] This point was made by J. MARSCHAK (1955) in his first paper on teams, "Elements for a Theory of Teams." I shall follow his good precedent.

From the standpoint of this chapter, then, the distinction between individual and organization will not be very important. A more significant taxonomy of theories of rational behavior, for our purposes, differentiates them by the assumptions they make about the "givens" – the given goals and given conditions. Particularly important is the distinction between those theories that locate all the conditions and constraints in the environment, outside the skin of the rational actor, and those theories that postulate important constraints arising from the limitations of the actor himself as an information processor. Theories that incorporate constraints on the information-processing capacities of the actor may be called *theories of bounded rationality*.

Rationality in the classical theory of the firm. The classical theory of the firm in its simplest form provides a useful standard for comparing and differentiating theories of rationality. In the theory of the firm, the given objective is to maximize profits, where profit is defined as the difference between gross receipts from sales and cost of production. The given conditions are two in number:

(I) *the demand function*: the quantity demanded is a function of price:

$$q_d = D(p), \text{ or } p = D^{-1}(q_d). \tag{1}$$

Since gross receipts equal price times quantity, the demand function determines gross receipts:

$$R = pq_d. \tag{2}$$

(II) *the cost function*: the cost of production is a function of the quantity produced:

$$C = C(q_s). \tag{3}$$

If the quantity produced equals the quantity demanded,

$$q_s = q_d, \tag{4}$$

then the profit, to be maximized, is simply the difference between gross receipts and the cost of production:

$$\text{Profit} = R - C = pq - C(q), \tag{5}$$

and, under appropriate assumptions regarding differentiability, we will

WHARTON REPROGRAPHICS

have for the maximum profit:

$$\mathrm{d}(R-C)/\mathrm{d}q = p+q\mathrm{d}(D^{-1}(q))/\mathrm{d}q-\mathrm{d}C(q)/\mathrm{d}q = 0\,. \tag{6}$$

The constraints in this theory, the demand and cost functions, D and C, are both located in the actor's environment. He is assumed to find the solution of equation (6). To do this, he must have perfect knowledge of these constraints, and must be able to perform the necessary calculations – to set the derivative of profit with respect to quantity equal to zero and to solve the resulting algebraic equation.

The limits of rationality. Theories of bounded rationality can be constructed by modifying these assumptions in a variety of ways. *Risk and uncertainty* can be introduced into the demand function, the cost function, or both. For example, certain parameters of one or both of these functions can be assumed to be random variables with known distributions. Then the assumption of the actor's perfect knowledge of these functions has been replaced by the assumption that he has perfect knowledge of their distributions. This change in assumptions may, in turn, make it easier or more difficult to carry out the calculations for finding the optimum – usually it becomes much more difficult than in the corresponding case of certainty.

Another way in which rationality can be bounded is by assuming that the actor has only *incomplete information about alternatives.* Fewer models have been constructed to deal with this situation than with the situation in which he has incomplete information about consequences. However, in certain search models it is assumed that the actor knows the probability distribution of profits in a population of possible alternative actions. Specific actions become available to him – say, by random sampling from this population – as a function of the amount of resources he devotes to search. His task is to find the alternative that maximizes his expected profit net of the search cost. In this class of models, selecting the best alternative from among those already discovered is assumed to be a trivial problem; the decision question has been switched to the question of how much of the actor's resources should be allocated to search.[2]

---

[2] For an example, see STIGLER (1961). Theories of the allocation of resources to search can also be constructed to deal with incomplete information about consequences. Sequential sampling theory falls into this category, for it answers the question: shall I make a decision now, or wait until I have gathered additional information? The question is answered by comparing the incremental cost of enlarging the sample with the expected gain through the resulting average improvement in the decision.

Finally, rationality can be bounded by assuming *complexity* in the cost function or other environmental constraints so great as to prevent the actor from calculating the best course of action. Limits on rationality stemming from this source have not been prominent in classical theories of rational behavior. However, in numerical analysis, the theory of approximation provides analogues, for it is concerned with the rate at which an approximation can be expected to improve as a function of amount of computational effort. By introducing explicitly into that theory the cost of computational effort, it can be transformed into a theory of optimal approximation.

Alternatives to the classical goals. The classical theory can be modified not only by altering the nature of the conditions and constraints, but also by altering the nature of the given goals. Some modern theories of the firm depart from the classical theory, not along any of the dimensions mentioned above, but by postulating different goals from the classical goal of profit maximization.

BAUMOL (1959, pp. 45–53), for example, has developed a model in which the firm maximizes sales subject to the constraint that profit should not be less than a specified "satisfactory" level. According to this theory of Baumol, equation (6) in the classical model should be replaced by:

$$\mathrm{d}R/\mathrm{d}q = p+q\mathrm{d}(D^{-1}(q))/\mathrm{d}q = 0\,, \tag{6'}$$

subject to the constraint that

$$P = R-C \geq P^*\,. \tag{7}$$

It may be observed that the informational and computational requirements for applying Baumol's theory to concrete situations are not very different from the requirements of the classical model.

This essay will not be concerned with variants of the theory of rationality that assume goals different from profit or utility maximization, except to the extent that there is significant interaction between the assumptions about goals and the assumptions about conditions and constraints. We shall see, however, that this is a very important exception. In actual fact, most of the variants of the theory that make significant modifications in the assumptions about conditions and constraints also call for assumptions about goals that are different from the classical assumptions of profit or utility maximization. The reasons for this interaction will appear as we proceed.

2. Approaches to Rational Choice in Chess

A number of the persons who have engaged in research on rational decision-making have taken the game of chess as a microcosm that mirrors interesting properties of decision-making situations in the real world. The research on rational choice in chess provides some useful illustrations of alternative approaches to rationality.

The problem confronting a chess player whose turn it is to move can be interpreted in either of two ways. First, it can be interpreted as a problem of finding a good (or the best) strategy – where "strategy" means a conditional sequence of moves, defining what move will be made at each successive stage after each possible response of the opponent.

Second, the problem can be interpreted as one of finding a set of accurate evaluations for the alternative moves immediately before the player.

From a classical standpoint, these two problems are not distinguishable. If the player has unlimited computational power, it does not matter whether he selects a complete strategy for his future behavior in the game, or selects each of his moves, one at a time, when it is his turn to play. For the way in which he goes about evaluating the next move is by constructing alternative complete strategies for the entire future play of the game, and selecting the one that promises the best return (i.e., the best return under the assumption that the opponent will also do *his* best to win). This is the approach taken in the von Neumann-Morgenstern theory of games (VON NEUMANN and MORGENSTERN (1953)).

The game-theoretical definition of rationality in chess. As von Neumann and Morgenstern observed, chess is a trivial game. "... if the theory of Chess (i.e., the complete tree of possible games) were really fully known there would be nothing left to play" (*ibid.*, p. 125). Each terminus of the tree of possible games represents a win, loss, or draw for White. Moving backward one branch on the tree, the player whose move it is at that branch can examine the termini to which it could lead by his choice of move, and can choose the move having the preferred terminus. The value of that terminus becomes, then, the value of the branch that leads to it. Working backward in this way, a value – win, lose, or draw for White – can be assigned to each position, and ultimately to each of the initial legal moves for White. Now each player can specify an optimal strategy – a strategy that will guarantee him at least as good an outcome as any other – by specifying which move he would select at each branch point in the tree whenever it is his move.

Unfortunately, as von Neumann and Morgenstern also observed, the triviality of chess offers no practical help to a player in actually choosing a move. "But our proof, which guarantees the validity of one (and only one) of these three alternatives [that the game must have the value of win lose or draw for White], gives no practically usable method to determine the true one. This relative, human difficulty necessitates the use of those incomplete, heuristic methods of playing, which constitute 'good' Chess; and without it there would be no element of 'struggle' and 'surprise' in that game" (*ibid.*).

What "impracticality" means becomes more vivid when we calculate how much search would be involved in finding the game-theoretically correct strategy in chess. On the average, at any given position in a game of chess, there are about 30 legal moves – in round numbers, for a move and its replies, an average of about 10^3 continuations. Forty moves would be a not unreasonable estimate of the average length of a game. Then there would be perhaps 10^{120} possible games of chess. Obviously the exact number does not matter: a number like 10^{40} would be less spectacular, but quite large enough to support the conclusions of the present argument.

Studies of the decision-making of chess players indicate strongly that strong players seldom look at as many as one hundred possibilities – that is one hundred continuations from the given position – in selecting a move or strategy. One hundred is a reasonably large number, by some standards, but somewhat smaller than 10^{120}! Chess players do not consider all possible strategies and pick the best, but generate and examine a rather small number, making a choice as soon as they discover one that they regard as satisfactory (see DE GROOT (1965)).

Before we consider in detail how they do it, let us return to the classical model and ask whether there is any way in which we could make it relevant to the practical choice problem, taking account of the size of the problem space, in a game like chess. One possible way would be to replace the actual problem space with a very much smaller space that approximates the actual one in some appropriate sense, and then apply the classical theory to the smaller approximate space.

This approach was taken in some of the early computer programs for playing chess and checkers. In the Los Alamos program, for example, the computer generated all legal moves, all legal replies to each, and so on, two moves deep. Each of the terminal positions thus generated (about a million in a two-move analysis) was evaluated, and the minimax procedure applied, working backwards, to find the best first move. Thus, a space of about

10^6 elements was substituted for the space of 10^{120} elements that represents the "real" world of chess.

The scheme was approximate, because the actual chess values of the million terminal positions were not known, and could not be known accurately without returning to the space of 10^{120} elements – that is, returning to the game-theoretical analysis of the full game. In place of these unknown true values, approximate values were computed, using rules of thumb that are commonly employed by chess players – conventional numerical values for the pieces, and measures of mobility. Thus, the approximate scheme was not guaranteed to select the objectively best move, but only the move leading to the positions that appeared best, in terms of these heuristic criteria, after an analysis two moves deep. Experience indicates that it is not possible to make such approximate evaluations accurately enough to enable the program to play good chess. The optimal decision in the approximated world is not necessarily even a good decision in the real world.

Satisficing processes in chess thinking. Chess programs now exist that take the alternative course, trying to emulate the human chess player in looking at only a very few continuations. The effectiveness of such a scheme depends critically on three components: the *move generators*, processes that select the continuations to be explored; the *evaluators*, processes that determine how good each continuation is; and the *stop rules*, criteria that determine when the search should be terminated and a move selected.

By scanning a chess position, features of the position can be detected that suggest appropriate moves. To take an extreme case, suppose a chess player discovers, when it is his move, that one of his Pawns attacks the opponent's Queen. Obviously, the capture of the Queen by the Pawn is one move that deserves consideration. It may turn out to be a poor move – another piece will checkmate him, say, if he captures the Queen – but its superficial merits are obvious, and its deficiencies can only be detected by considering it and evaluating it dynamically. A simple process that would generate this move, and others like it, would consist in determining which of the opponent's pieces were attacked by a piece of lesser value, or were undefended and attacked by any piece. Thus, a suitable set of move-generating processes might identify for further analysis all or most of the moves deserving serious consideration. If the generators were ordered appropriately, they might usually identify first the most promising moves, then the ones slightly less promising, and so on.

Possible moves, produced by the move generators, can be evaluated by a combination of static and dynamic criteria. Static criteria are features of the position, or differences between successive positions. Thus, one of the important static evaluators used by all chess players is the piece count: each piece is assigned a conventional value (say, Pawn = 1, Knight and Bishop = 3, Rook = 5, Queen = 9), and the sums of the values for the two players are compared. In general, if the piece count of one player exceeds that of the other by more than one point (or even, in many cases, by a single point), the player with the higher count can find a winning continuation *unless* the balance is very quickly redressed by a sequence of forceful moves. (Thus, it does not matter being 5 points down if you can capture the opponent's Queen on the next move without further reprisals.)

The short-run tactical considerations are handled by carrying out dynamic analysis of plausible continuations until a position is reached that is sufficiently quiet or "dead" that it can safely be evaluated by means of the static evaluators. These static evaluators are then propagated backwards to the move under consideration by the familiar minimax procedure.

Two kinds of stop rules are needed in a program having this structure: rules to stop exploration at dead positions that can be evaluated statically, and rules to stop the entire process and select a move when a satisfactory one has been found. The former class of stop rules has already been discussed; the latter needs to be examined more closely. If the alternatives in a choice situation are not given, but have to be discovered or invented, and if the number of possible alternatives is very large, then a choice has to be made before all or most of them have been looked at. It was precisely this difficulty in the classical requirement of comparing all alternatives that led to the approach described here. But if all alternatives are not to be examined, some criterion must be used to determine that an adequate, or satisfactory, one has been found. In the psychological literature, criteria that perform this function in decision processes are called *aspiration levels*. The Scottish word "satisficing" (= satisfying) has been revived to denote problem solving and decision making that sets an aspiration level, searches until an alternative is found that is satisfactory by the aspiration level criterion, and selects that alternative (SIMON (1957), Part IV).

In satisficing procedures, the existence of a satisfactory alternative is made likely by dynamic mechanisms that adjust the aspiration levels to reality on the basis of information about the environment. Thus, in a chess-playing program, the initial aspiration level can be set (preferably with a little upward bias) on the basis of a static evaluation of the position. As alternative moves are considered and evaluated by dynamic and static analysis, the evaluation of the position can gradually be reduced until the best move discovered so far reaches or exceeds in value the aspiration level.

The limits of rationality in chess. In the introductory section of this paper, three limits on perfect rationality were listed: uncertainty about the consequences that would follow from each alternative, incomplete information about the set of alternatives, and complexity preventing the necessary computations from being carried out. Chess illustrates how, in real world problem-solving situations, these three categories tend to merge.

If we describe the chess player as choosing a *strategy*, then his difficulty in behaving rationally – and the impossibility of his behaving as game theory says he should – resides in the fact that he has incomplete information as to what alternatives (strategies) are open to him. He has time to discover only a minute fraction of these strategies, and to specify the ones he discovers only incompletely.

Alternatively, if we describe the chess player as choosing a *move*, his difficulty in behaving rationally lies in the fact that he has only rough information about the consequences of adopting each of the alternatives (moves) that is open to him. It would not be impossible for him to generate the whole set of his legal moves, for they seldom number more than about thirty. However, he can evaluate them, even approximately, only by carrying out further analysis through the immense, branching, move tree. Since only a limited amount of processing time is available for the evaluation, he must allocate the time among the alternative moves. The practical facts of the matter are that it is usually better to generate only a few of the entire set of legal moves, evaluating these rather thoroughly, than it is to generate all of them, evaluating them superficially. Hence the good chess player does not examine all the moves open to him, but only a small fraction of them. (Data presented by DE GROOT (1965) suggest that typically a half dozen to a dozen of a set of thirty legal moves may be generated and explored by the chess player.)

From still a third standpoint, the chess player's difficulty in behaving rationally has nothing to do with uncertainty – whether of consequences or alternatives – but is a matter of complexity. For there is no risk or uncertainty, in the sense in which those terms are used in economics or statistical decision theory, in the game of chess. As von Neumann and Morgenstern observe, it is a game of perfect information. No probabilities of future events need enter the calculations, and no contingencies, in a statistical sense, arise.

From a game-theoretical standpoint, the presence of the opponent does not introduce contingencies. The opponent can always be counted on to do

his worst. The point becomes clear if we replace the task of playing chess with the task of proving theorems. In the latter task, there is no opponent. Nor are there contingencies: the true and the derivable theorems reside eternally in Plato's heaven. Rationality in theorem proving is a problem only because the maze of possible proof paths is vast and complex.

What we refer to as "uncertainty" in chess or theorem proving, therefore, is uncertainty introduced into a perfectly certain environment by inability – computational inability – to ascertain the structure of that environment. But the result of the uncertainty, whatever its source, is the same: approximation must replace exactness in reaching a decision. In particular, when the uncertainty takes the form of an unwieldy problem space to be explored, the problem-solving process must incorporate mechanisms for determining when the search or evaluation will stop and an alternative will be chosen.

Satisficing and optimizing. The terms satisficing and optimizing, which we have already introduced, are labels for two broad approaches to rational behavior in situations where complexity and uncertainty make global rationality impossible. In these situations, optimization becomes approximate optimization – the description of the real-world situation is radically simplified until reduced to a degree of complication that the decision maker can handle. Satisficing approaches seek this simplification in a somewhat different direction, retaining more of the detail of the real-world situation, but settling for a satisfactory, rather than an approximate-best, decision. One cannot predict in general which approach will lead to the better decisions as measured by their real-world consequences. In chess at least, good players have clearly found satisficing more useful than approximating-and-optimizing.

A satisficing decision procedure can often be turned into a procedure for optimizing by introducing a rule for optimal amount of search, or, what amounts to the same thing, a rule for fixing the aspiration level optimally. Thus, the aspiration level in chess might be adjusted, dynamically, to such a level that the expected improvement in the move chosen, per minute of additional search, would just balance the incremental cost of the search.

Although such a translation is formally possible, to carry it out in practice requires additional information and assumptions beyond those needed for satisficing. First, the values of alternatives must be measured in units comparable with the units for measuring search cost, in order to permit comparison at the margins. Second, the marginal productivity of search – the expected increase in the value per unit of search time – must be estimated on some basis or other. If one were designing a chess-playing program, it is doubtful whether effort spent in attempting to imbed the program in such a dynamic optimizing framework would be nearly as worthwhile as equivalent effort spent in improving the selectivity of the program's move-generating and move-evaluating heuristics.

Another quite different translation between optimizing and satisficing schemes has also been suggested from time to time. A chess program of the "classical" type, which makes optimal decisions in an approximated world, can be regarded as a particular kind of satisficing program, in which "satisfactory" is defined by the approximating procedure that is used. Hence, it is difficult to draw a formal distinction between optimizing and satisficing procedures that is so iron-clad as to prevent either from being reinterpreted in the frame of the other. The practical difference, however – the difference in emphasis that results from adopting one viewpoint or the other – is often very great.

In research on optimizing procedures, considerable attention has been paid to the formal properties of the evaluation functions, to the existence and efficiency of procedures for computing the optimum, and to procedures for reducing uncertainty (e.g., forecasting methods). The nature of the approximations that are necessary to cast real-world problems into forms suitable for optimization, and the means for choosing among alternative approximations, have been less fully and less systematically studied. Much effort, for example, has gone into the discovery of efficient algorithms for solving linear programming problems. Finding an appropriate way of formulating a concrete real-world decision problem as a linear-programming problem remains largely an art.[3]

Research on satisficing procedures has focussed primarily on the efficiency of search – on the nature of the heuristic methods that enable the rare solutions in enormous spaces of possibilities to be sought and found with moderate amounts of search effort. Since moderate changes in heuristics often make order-of-magnitude changes in search effectiveness, highly accurate means for assessing the quality of solutions or the effort required to find them may be relatively unimportant. It probably does not require delicate evaluation functions or stop rules to change a duffer's chess play to a reasonably effective move-choosing program.

---

[3] The work of A. CHARNES and W. W. COOPER (1961) is full of sophisticated examples of this art. See, for instances, Appendix B and Chapter 11 of Volume I.

3. Bounded Rationality in Design

The engineering activities usually called "design" have not been much discussed under the heading of rational decision-making. The reason for this should be clear from the foregoing discussion: classical decision theory has been concerned with choice among *given* alternatives; design is concerned with the discovery and elaboration of alternatives. Our exploration of the microcosm of chess has indicated, however, how the theory of design can be assimilated to a satisficing theory of rational choice. Let me spell the point out a little more fully.

Consider that interpretation of chess which views the task as one of choosing a strategy, and not just a single move. Specifically, consider a situation where a player is searching for a combination (a strategy) that will definitely checkmate his opponent, even though it may require sacrifices of pieces along the way. A chess player will ordinarily not enter into such a course of action unless he can see it through to the end – unless he can *design*, that is, a water-tight mating combination.

As we have seen already, the evaluations and comparisons that take place during this design process are not, in general, comparisons among complete designs. Evaluations take place, first of all, to guide the search – the elaboration of the design itself. They provide the basis for decisions that the design should be elaborated in one direction rather than another. Complete designs (in this case, mating combinations), when they are finally arrived at, are not generally evaluated by comparing them with alternative designs, but by comparing them with standards defined by aspiration levels. In the chess situation, as soon as the player discovers a strategy that guarantees a checkmate, he adopts it. He does not look for all possible checkmating strategies and adopt the best (H. A. SIMON and P. A. SIMON (1962)).

In the design of complex objects – a bridge, say, or an airplane – the process has an even more involved search structure. Here, the early stages of search take place in highly simplified spaces that abstract most of the detail from the real-world problem, leaving only its most important elements in summarized form. When a plan, a schematized and aggregated design, has been elaborated in the planning space, the detail of the problem can be reintroduced, and the plan used as a guide in the search for a complete design.

More than two spaces may be used, of course; there may be a whole hierarchy of planning spaces, leading from a highly abstract and global design to successive specification of detail. At each of these levels of abstraction, the design process, too, may be differently structured. Since the more abstract spaces tend to be "smoother," it is often possible to use optimization models for planning purposes, reverting to satisficing search models to fill in the detail of the design. Thus, linear programming or dynamic programming may be used for general planning of factory operations, while more heuristic techniques are used for scheduling of individual jobs. In other situations, the overall design process may employ satisficing search procedures, while optimizing techniques may be used to set parameters once the general design has been fixed.[4]

4. Bounded Rationality in Management Science

Most of the formal techniques that constitute the technical backbone of management science and operations research are procedures for finding the best of a set of alternatives in terms of some criterion – that is, they fall in our category of "classical" procedures. Linear and dynamic programming are among the most powerful of these techniques. The dominant approach to problems in this sphere has been to simplify the real-world problems to the point where the formal optimizing models can be used as approximations.

Some industrial problems of a combinatorial sort have not yielded easily to this approach. Typically, the recalcitrant problems involve integer solutions, or, what usually amounts to the same thing, the consideration of possible permutations and combinations of a substantial number of elements. Warehouse location is a problem of this kind. The task is to "determine the geographical pattern of warehouse locations which will be most profitable to the company by equating the marginal cost of warehouse operation with the transportation cost savings and incremental profits resulting from more rapid delivery" (KUEHN and HAMBURGER (1963), p. 643).

A heuristic program devised by KUEHN and HAMBURGER (1963) for locating warehouses has two parts: "(1) the main program, which locates warehouses one at a time until no additional warehouses can be added to the distribution network without increasing total costs, and (2) the bump and shift routine, ..., which attempts to modify solutions ... by evaluating the profit implications of dropping individual warehouses or of shifting them from one location to another" (*Ibid.*, p. 645).

---

[4] Some modern semi-automated procedures for the design of chemical processing plants proceed from heuristic techniques for selecting the unit operations and their flow, then employ linear programming to determine the parameters of the system so specified.

This program fits our earlier characterization of design procedures. A possible plan is gradually built up, step by step, through a search procedure, and then possible local modifications are investigated before the final plan is settled upon. In building up the initial plan, locations are tried that are near concentrations of demand, adding at each step the warehouse that produces the greatest cost savings for the entire system. Only a fraction of the possible warehouse sites, which preliminary screening selects as "promising," are evaluated in detail at each stage. Finally, a so-called "bump-shift" routine modifies the programs tentatively arrived at by (1) eliminating warehouses no longer economical because new warehouses have been introduced at later steps of the program, (2) considering shifting warehouses to alternative sites within their territories. The flow diagram of the warehouse location programs, which will serve to illustrate the typical structure of heuristic programs when they are formalized, is shown in Fig. 1.

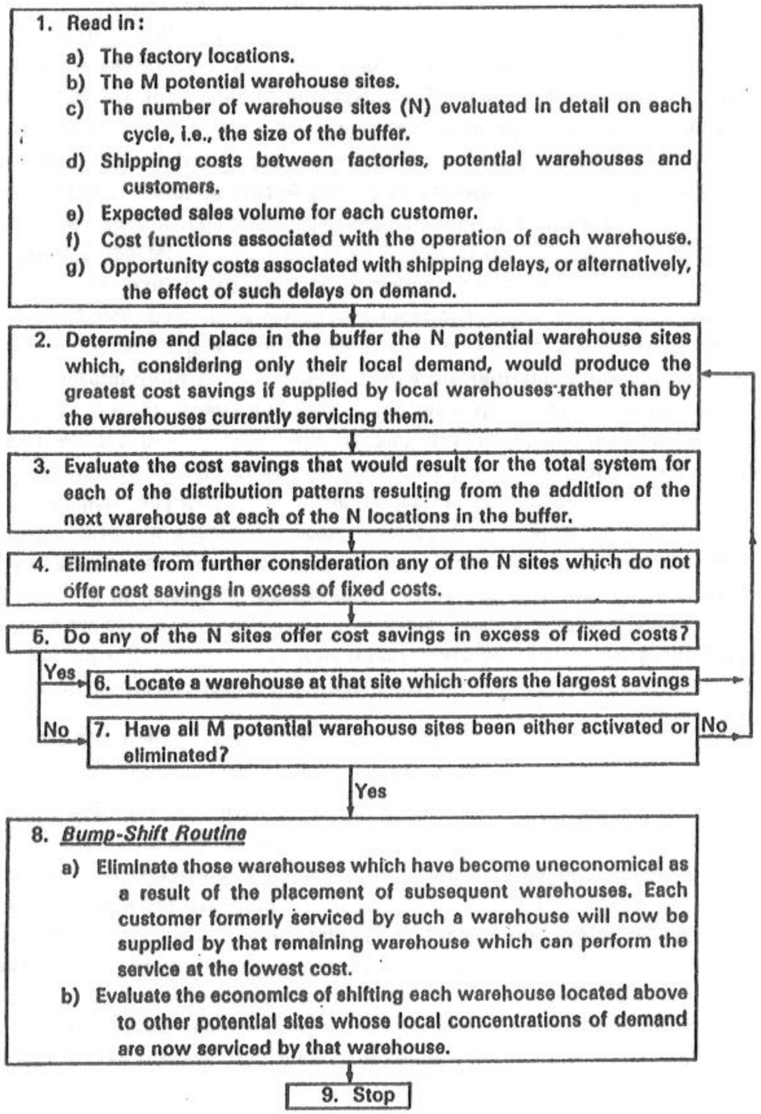
Kuehn and Hamburger have carried out some detailed comparisons of the heuristic program with optimizing techniques. They conclude that "in theory, a linear programming approach ... could be used to solve the problem. In practice, however, the size and nonlinearities involved in many problems are such that application is not currently feasible" (*ibid.*, p. 658). They attribute the superior performance of the heuristic program to two main causes: "(1) computational simplicity, which results in substantial reductions in solution times and permits the treatment of large-scale problems, and (2) flexibility with respect to the underlying cost functions, eliminating the need for restrictive assumptions" (*ibid.*, p. 656).

Perhaps the technique most widely used in management science to deal with situations too complex for the application of known optimization methods is simulation. The amount of detail incorporated in the simulation of a large system is limited only by computational feasibility. On the other hand, simulation, unaided by other formal tools of analysis, provides no direct means for discovering and evaluating alternative plans of action. In simulation, the trial and error is supplied by the human investigators rather than by the technique of analysis itself (see FORRESTER (1961)).

5. Conclusion

The theory of rational decision has undergone extremely rapid development in the past thirty years. A considerable part of the impetus for this development came, during and since World War II, from the attempt to use formal decision procedures in actual real-world situations of considerable

A HEURISTIC PROGRAM FOR LOCATING WAREHOUSES



1. Read in:
   - a) The factory locations.
   - b) The M potential warehouse sites.
   - c) The number of warehouse sites (N) evaluated in detail on each cycle, i.e., the size of the buffer.
   - d) Shipping costs between factories, potential warehouses and customers.
   - e) Expected sales volume for each customer.
   - f) Cost functions associated with the operation of each warehouse.
   - g) Opportunity costs associated with shipping delays, or alternatively, the effect of such delays on demand.
2. Determine and place in the buffer the N potential warehouse sites which, considering only their local demand, would produce the greatest cost savings if supplied by local warehouses rather than by the warehouses currently servicing them.
3. Evaluate the cost savings that would result for the total system for each of the distribution patterns resulting from the addition of the next warehouse at each of the N locations in the buffer.
4. Eliminate from further consideration any of the N sites which do not offer cost savings in excess of fixed costs.
5. Do any of the N sites offer cost savings in excess of fixed costs?
6. Locate a warehouse at that site which offers the largest savings
7. Have all M potential warehouse sites been either activated or eliminated?
8. *Bump-Shift Routine*
   - a) Eliminate those warehouses which have become uneconomical as a result of the placement of subsequent warehouses. Each customer formerly serviced by such a warehouse will now be supplied by that remaining warehouse which can perform the service at the lowest cost.
   - b) Evaluate the economics of shifting each warehouse located above to other potential sites whose local concentrations of demand are now serviced by that warehouse.
9. Stop

Fig. 1. Flow diagram

complexity. To deal with this complexity the formal models have grown in power and sophistication. But complexity has also stimulated the development of new kinds of models of rational decision that take special account of the very limited information-gathering and computing capacity of human beings and their associated computers.

One response to the concern with uncertainty, with the difficulties of discovering or designing alternatives, and with computational complexity has been to introduce search and information transmission processes explicitly into the models. Another (not exclusive) response has been to replace optimization criteria with criteria of satisfactory performance. The satisficing approach has been most often employed in models where "heuristic" or trial-and-error methods are used to aid the search for plausible alternatives.

As a result of all these developments, the decision maker today, in business, government, universities, has available to him an unprecedented collection of models and computational tools to aid him in his decision-making processes. Whatever the compromises he must make with reality in order to comprehend and cope with it, these tools make substantially more tractable the task of matching man's bounded capabilities with the difficulty of his problems.

References

BAUMOL, W. J. (1959), *Business Behavior, Value and Growth*, Macmillan, New York, pp. 45–53.

CHAPANIS, N. P. and J. A. CHAPANIS (1964), "Cognitive Dissonance: Five Years Later," *Psychological Bulletin*, *61*, 1023.

CHARNES, A. and W. W. COOPER (1961), *Management Models and Industrial Applications of Linear Programming*, Wiley, New York, (2 volumes).

DE GROOT, A. (1965), *Thought and Choice in Chess*, Mouton, The Hague.

FORRESTER, J. W. (1961), *Industrial Dynamics*, M.I.T. Press, Cambridge.

KUEHN, A. A. and M. J. HAMBURGER (1963), "A Heuristic Program for Locating Warehouses," *Management Science*, *9*, 643-666.

MARSCHAK, J. (1955), "Elements for a Theory of Teams," *Management Science*, *1*, 127–137

SIMON, H. A. (1957), Part IV in *Models of Man*, Wiley, New York, pp. 196–279.

SIMON, H. A. and P. A. SIMON (1962), "Trial and Error Search in Solving Difficult Problems," *Behavioral Science*, *7*, 425–429.

STIGLER, G. J. (1961), "The Economics of Information," *Journal of Political Economy*, *69*, 213–225.

VON NEUMANN, J. and O. MORGENSTERN (1953), *Theory of Games and Economic Behavior*, (3rd ed.), Princeton University Press, Princeton, pp. 125.

SELECTED TOPICS IN INDETERMINISTIC SYSTEMS

Aron Katsenelinboigen

THE SYSTEMS INQUIRY SERIES
Published by Intersystems Publications

10 A

WHARTON REPROGRAPHICS — 2-3

183

7. CHESS AS A MULTI-STAGE GAME

Contents

Introductory remarks 100

Chapter 1. The set up of the chess game

§ 1. The place of chess 105
§ 2. The function and the structure of the game 107
§ 3. The rules and the operator of the game 110
§ 4. The genesis of the game 115

Chapter 2. The game at the global level

§ 1. Optimization vs. chance. An inexact problem 118
§ 2. Algorithms of inexact problems - general description 121
§ 3. Reactive methods 123
§ 4. Search solutions 128

Chapter 3. Evaluation of a state: parameters of integration

§ 1. Analytical representation of the holistic effect 132
§ 2. Evaluation of material 134
§ 3. The structure of positional parameters 140

Chapter 4. Class of local problems and stages of the game

§ 1. Classification of search problems 149
§ 2. Non-goal directed protractive problems 151
§ 3. Goal-directed protractive problems 154
§ 4. Com-positional problems 157
§ 5. Stages of the game 160

Chapter 5. Styles of the game

§ 1. The subjective and the objective in chess: aesthetic and scientific methods 165
§ 2. A sacrifice in chess 168
§ 3. Logical and historical development of styles of play 171

Conclusion 177

Introductory remarks

Methods of formulating and solving a problem are many and diverse. Oftentimes, considerable ingenuity is required in order to determine which information belongs in the formulation of the problem and which in its solution.

I wanted to remind the reader of this well known phenomenon in order to emphasize the peculiar nature of the subject investigated in the present section of the book. Unlike general philosophical problems concerned with the fundamental questions of the 'essence of the world','the meaning of life', etc. which still remain open, chess represents a problem that has already been formulated in a very precise fashion: our task is to solve it. Although the problem of having to set local goals does arise in the course of advancing toward the global goal of the game, the latter is given to us beforehand.

Naturally, our analysis is greatly simplified when the problem is formulated in a complete and general form as is the case in chess. Clear cut rules and a certain amount of idealization allows us to concentrate on the inner structure of a system such as chess. Although chess is limited because of its rigid structure it is still very useful as a model in exploring certain problems in different areas of arts and sciences.

For many writers, the game of chess was an allegory of human relationships (for instance, Nabokov's novel Luzhin's Defense) or human psychology (like Poe's story The Murders in the Rue Morgue). Scholars and artists in many different fields though chess to be a good analogy to their craft. Direct associations between chess, especially its gambit openings and musical composition, its beginning in particular were noted by L. Rowell.

> He wrote: "One could easily write a history of music based on beginning mannerisms and gambits. In fact the analogy to the game of chess is a very good one: just as in chess there are certain combinations of opening moves (gambits) that implement the strategic principles of the opening game, namely (1) the rapid development of pieces and (2) control of the center of the board, in music there are various gestures which are designed to achieve some important strategic objectives of what we might call the "opening game of music."[1]

De Saussure compared the game of chess with language both in its formulation and methods of solution. For de Saussure the most valuable idea he extracted from chess was the precise understanding of the notion of the value of a piece; he emphasized the use of this idea in linguistics to help determine the significance of words in a sentence. De

---

[1] Rowell, L., "The Creation of Audible Time", The Study of Time IY, Papers from the Fourth Conference of the International Society for the Study of Time, Alpbach-Austria, ed. by J. Fraser, etc., New York: Springer-Verlag, pp. 198-210.

Saussure's observations on the connections between language and chess are so original that I want to quote them in full.[2]

Scholars did not employ chess merely as an analogy. Various methods of playing the game were used directly to solve some very practical problems and economic ones in particular[3]. The range of application of these methods is quite wide.

---

[2] "In internal linguistics the picture differs completely. Just any arrangement will not do. Language is a system that has its own arrangement. Comparison with chess will bring out the point. In chess, what is external can be separated relatively easily from what is internal. The fact that the game passed from Persia to Europe is external; against that, everything having to do with its system and rules is internal. If I use ivory chessmen instead of wooden ones, the change has no effect on the system; but if I decrease or increase the number of chessmen, this change has a profound effect on the "grammar" of the game. One must always distinguish between what is internal and what is external. In each instance one can determine the nature of the phenomenon by applying this rule: everything that changes the system in any way is internal." (pp. 22-23).

"But of all comparisons that might be imagined, the most fruitful is the one that might be drawn between the functioning of language and a game of chess. In both instances we are confronted with a system of values and their observable modifications. A game of chess is like an artificial realization of what language offers in a natural form.

Let us examine the matter more carefully. First, a state of the set of chessman corresponds closely to a state of language. The respective value of the pieces depends on their position on the chessboard just as each linguistic term derives its value from its opposition to all the other terms.

"In the second place, the system is always momentary; it varies from one position to the next. It is also true that values depend above all else on an unchangeable convention, the set of rules that exist before a game begins and persists after each move. Rules that are agreed upon once and for all exist in language too; they are the constant principles of semiology.

Finally, to pass from one state of equilibrium to the next, or according to our terminology-from one synchrony to the next, only one chess piece has to be moved; there is no general rummage. here we have the counterpart of the diachronic phenomenon with all its peculiarities. In fact:

(a) In each play only one chess piece is moved; in the same way in language, changes affect only isolated elements.

(b) In spite of that, the move has a repercussion on the whole system; it is impossible for the player to foresee exactly the extent of the effect. Resulting changes of value will be, according to the circumstances, either nil, very serious, or of average importance. A certain move can revolutionize the whole game and even affect pieces that are not immediately involved. We have just seen that exactly the same holds for language.

(c) In chess, each move is absolutely distinct from the preceding and the subsequent equilibrium. The change effected belongs to neither state: only states matter.

In a game of chess any particular position has the unique characteristic of being freed from all antecedent positions; the route used in arriving there makes absolutely no difference; one who has followed the entire match has no advantage over the curious party who comes up at a critical moment to inspect the state of the game; to describe this arrangement, it is perfectly useless to recall what had just happened ten seconds previously. All this is equally applicable to language and sharpens the radical distinction between diachrony and synchronic. Speaking operates only on a language state, and the changes that intervene between states have no place in either state.

For instance, in 1984 the well known Boston Consulting Group published a brochure entitled "Business Chess". It notes that despite chess' relative simplicity it can still serve as a model to help solve economic problems. This point is substantiated by the following observation:

> "...they [chess and economics] are systems of competition, and the competition involves indirect consequences."

Authors of the brochure provide several examples of chess being utilized in the world of business. These examples deal with various chess concepts such as gambits, establishing the value of the pieces under different circumstances, recognition of a position, various chess techniques elaborated over the course of time, and the need to learn. Jumping ahead a little I want to note that the authors of the brochure omit the use of prior experience in formulating and solving local problems, i.e. problems arising within the search tree.

Chess ideas were put to much wider use by the former world chess champion M. Botvinnik. In particular, Botvinnik's article in the newspaper "Pravda", March 30, 1981 described a program for scheduling the repair work of the Central Dispatching Office of the United Energy System of the USSR. This program was based on chess algorithm called "Pioneer 2" which he helped develop.

Naturally, chess methods found their immediate application in the field of artificial intelligence. H. Simon stressed the role of chess viewing it as a

> "...microcosm that mirrors interesting properties of decision-making situations in the real world."[4]

I only want to note that many scholars consider chess a credible source for studying artificial intelligence because the game, due to its clear cut and rigid rules, yields to formal representation. Chess possesses a symbolic language which can be used to define the game. Note that this language can be understood by people with different native tongues. In recent times this language not only acquired a chess notation which refers to individual pieces, their coordinates and moves, but it also incorporates the analysis of how the game is actually proceeding.[5] The symbol + means advantage for white; Δ - with an idea; ˙ - the only move, etc. Positional parameters are also given a symbolic representation. ∞ denotes connected pawns, ∘...∘ stands for isolated pawns, etc.

But the major advantage of chess in the field of artificial intelligence. It seems to me, at least the following three features make chess so eligible in this area.

The first is due to the fact that chess represents a model of system's development which does not yield (perhaps even theoretically) to a constructive algorithm which would link completely (in a sense of complete certainty of the outcome) a given step with the final outcome.

The other peculiar aspect of chess is that it can be viewed as a multi-stage game. A stage, from an operational point of view, represents a certain interval of the overall process which employes a particular method of operation. Thus, multi-stage processes give rise to the problem of finding an adequate set of methods for each stage and their subsequent integration. Chess is a treasure chest of experience when it comes to multi-stage processes, a fact which is evident in the advanced theory of openings, middle game, and endgame.

The third feature pertains to the positional method. Of all the various methods used in this ancient game the positional method practically owes its advancement to the twentieth century. This very original method incorporates unique ideas regarding analytical

---

At only one point is the comparison weak: the chess player intends to bring about a shift and thereby to exert an action on the system, whereas language premeditates nothing. The pieces of language are shifted-or rather modified-spontaneously and fortuitously. The umlaut of Hande for hanti and Gaste for gasti (see p. 83) produced a new system for forming the plural but also gave rise to verbal forms like tragt from tragit, etc. In order to make the game of chess seem at every point like the functioning of language, we would have to imagine an unconscious or unintelligent player. This sole difference, however, makes the comparison even more instructive by showing the absolute necessity of making a distinction between the two classes of phenomena in linguistics. For if diachronic facts cannot be reduced to the synchronic system which they condition when the change is intentional, all the more will they resist when they set a blind force against the organization of a system of signs." (pp. 88-89).
"Finally, not every idea touched upon in this chapter differs basically from what we have elsewhere called values. A new comparison with the set of chessmen will bring out this point (see pp. 88 ff.). Take a knight, for instance. By itself is it an element in the game? Certainly not, for by its material make-up--outside its square and the other conditions of the game--it means nothing to the player; it becomes a real, concrete element only when endowed with value and wedded to it Suppose that the piece happens to be destroyed or lost during a game. Can it be replaced by an equivalent piece? Certainly. Not only another knight can be declared identical provided the same value is attributed to it. We see then that in semiological systems like language, where elements hold each other in equilibrium in accordance with fixed rules, the notion of identity blends with that of value and vice versa." Saussure, F., Course in General Linguistics, New York: McGraw Hill Book Co., 1959, p.110.

[3] Botvinnik, M., "Decision Making and Computer", Advances in Computer Chess, 3, ed. by M. R. B. Clarke. Oxford: Pergamon Press, 1982, pp. 169-179; Nigan, R.; "Operations Research/Management Science and Chess", RCA Engineer, Vol. 29, No. 4, 1984, pp. 69-73.

[4] Simon, H., "Theories of Bounded Rationality", Models of Bounded Rationality, vol. 2. Cambridge: The MIT Press, 1982, p. 412.

[5] "Chess Informant", 1980.

evaluation of a position - evaluation which can be reduced to numbers. I know of no other area exhibiting the same precision when it comes to evaluating a state of the system. Axiology has not as of yet absorbed these revolutionary ideas into its arsenal. Nevertheless the role of positional method is enormous for it forms the foundation of the aesthetic method of judgement by means of which indeterministic processes permeating all different areas of human endeavor are implemented.

Finally, chess is highly useful in the field of artificial intelligence in that it yields to an experimental verification of various ideas. A famous Soviet mathematician A. Kronrod, one of the pioneers of computer chess in the Soviet Union, once wrote that the role of chess in artificial intelligence could be compared to that of a drosophila fly in genetics.[6] On the one hand, chess' peculiarity as a multi-stage game, its relative simplicity as compared to many other systems, and on the other, fast attainment of the final state[7] and statistics that can be gathered because we can always come back to the initial state, give us reasons to believe that chess can become a very useful experimental tool.

Now I want to say a few words about the structure of the present section.

It has already been said that the end goal in chess is given. Still a few comments regarding this aspect of the problem are in order and this constitutes the subject matter of the first section. It basically consists of comments regarding the problem of goal formulation. This approach is different from a more analytic one used in discussing the solution of the problem. The greater part of the chess section of the book deals with the latter topic. The section is divided into four chapters.

The division was based on the generality of the problem in question. This criterion gave rise to the following grouping: global problems pertaining to the overall flow of the game, local problems embracing a few moves, and problems occurring at intermediate levels - stages of the game.[8] Each of these subjects comprises a chapter. Importance of the problem of evaluating any given state of the game warranted a separate chapter.

---

[6] Kronrod, A. S., "Machine is Getting More Intelligent", Pravda, March 15, 1967.

[7] Constraints on the duration of the game is not without drawbacks. For instance, each side in a given game is represented by a single player. Real life development of an idea may require significantly more time than the life span of its inventor. Here a serious problem of ensuring proper transfer of power comes under consideration. Chess could perhaps mimic this aspect if something like a relay race was introduced with the game being handed from one player to another. Here participant selection would become an issue.

[8] Division of the field into a macro and a micro, as is the case with economics, seems inadequate for our discussion because certain intermediate stages significant in themselves are left unaccounted. Our case is closer to the classification of stages in planning where a distinction is made between strategic, mid-term, and tactical planning.

Chapter 1. The set up of the chess game

§1. The place of chess.

Let us start by immersing chess in a more general system.

By its nature chess belongs to the realm of art rather than science (taking the criterion of art to be the creation of models which do not necessarily reflect reality). What distinguishes these models from the scientific ones is that they do not require experimental verification to determine whether or not they correspond to some real phenomena outside of themselves.[9]

But if we view science as that sphere of human thought which produces objective knowledge, i.e. knowledge that can be passed on to other people who can use it with the same results, then chess should be part of sciences since methods used in the game allow for this kind of knowledge to be formed.

Therefore, chess occupies a very peculiar place in the system. It can be included among the arts as well as sciences. Moreover, although the formulation of the game places it in the category of art chess does reflect some very important aspects of reality. Judging by the methods used in the game chess can be classified as a science although an aesthetic method which prevails in the creation and perception of art plays an important role in chess.

So what is the role of chess as far as it serves these two dieties? Just like other games chess fulfills many functions: it is played for recreation, it is used to teach certain techniques of decision making, and to influence the behavior of people. All this is accomplished by replacing the "real" actions with their model simulation. Chess is rather peculiar in this respect. I shall explain what I mean by this statement later on.

Like many other games chess belongs to the class of games whose purpose is to simulate certain aspects of real life. It may seem at first glance that chess merely reflects real life conflicts, or to be more precise, it simulates military combat.

But this approach to chess and its connection with reality is rather limited and superficial. Thinking of chess as a constructive model capable as such (i.e. its specific methods of play ignored) of affecting interrelationships among people we begin to see that its connection with reality is much broader and deeper.

---

[9] This is precisely the criterion used in awarding the Nobel Prize. It is not an accident that the peace prize and the award for literature are placed in a different category. Nobel prize is not awarded for mathematics since according to this criterion mathematics belongs in the realm of art. (Bitter tongues say that Nobel excluded mathematics from the list of disciplines for which the prize is awarded because his wife's lover was a mathematician).

It seems to me that reducing chess to a game whose sole purpose is to simulate a conflict situation is an oversimplification. To clarify this point I want to call to the attention of the reader the following classification of relations, including conflicts, proposed by J. Gharajedaghi and based on the ideas of R.Ackoff.[10]

Table 1. Types of Interrelations

Means	Ends	
	Incompatible	Compatible
Compatible	Collaboration	Cooperation
Incompatible	Conflict	Competition

This taxonomy gives rise to four types of relations. Two of them correspond to conflict situations. What makes competition different from a conflict is that the goals of the participants are compatible while the means are incompatible. Here, the term goal stands for a noble goal such as the enjoyment of the game; the goal of winning merely represents means of achieving this end.

This idea was expressed independently and in different words by A. Koestler. He wrote:

> "Edward Lasker (namesake of the great Emanuel, and himself a grand master) wrote a revealing book with the title: Chess for Fun and Chess for Blood. But "fun" is the wrong word; what he meant was that the game of chess is the perfect paradigm for both the glory and the bloodiness of the human mind. On the one hand, an exercise in pure imagination happily married to logic, staged as a ballet of symbolic figures on a mosaic of sixty-four squares; on the other hand a gladiatorial contest. This dichotomy is perhaps the main secret of the game's astonishingly long history."[11]

Actually, there are some forms of chess where players relate to each other in ways other than purely competitive. I mean the 4-player version of chess which was played in India. There were two sets of pieces - green and white on one side and black and red on another. These two groups related to each other in a competitive mode. Within each of these groups relations were of a cooperative type since both the ends and the means of the participants were compatible.

---

[10] Gjarajedaghi,J., Toward a Systems Theory of Organization. Seaside: Intersystems Publications, 1985, pp. 41-42.

[11] Koestler, A., The Hill of Achilles. London: Hutchinson of London, 1974, p.196.

So according to the classification proposed above chess represents a game which simulates competition since the goals of the participants are compatible while the means are not.

§2. The function and the structure of the game.

Now, let us take a look at the final product of the game and its role in the fulfillment of the aforementioned functions.

Chess belongs to the class of games having a well defined final goal (unlike other games lacking such goals - for instance, a man running around playing with a ball). The goal in chess is to capture one very special material object - the opponent's king. Victory is characterized by the state of one single piece, the king. The state of the system as a whole is of no consequence at all. In other words, the price of a victory is irrelevant: you can lose almost all your pieces as long as you capture the enemy's king. So Pyrrhic victory which is so very dangerous for systems having no end goals (they develop continuously) is impossible in chess. Actually Pyrrhic victory has its significance for chess but only as far as intermediate stages of the game are concerned.

We can imagine a form of chess having a different objective with an overall state of the system being evaluated at the end of the game. Here, it becomes necessary to assign external values not only to the king but to other pieces as well. The more pieces are left on the board, both your own and your opponent's, the greater is your victory.

Now, on the structure of chess. There is a considerable variety of pieces in chess. They differ in the type of moves they are allowed to make.

This feature distinguishes chess from many other games where the number of objects manipulatable by the players is limited both in quantity (soccer, baseball) and quality (bowling, checkers). The physical weight of the piece and their volume are insignificant. Pieces operate within the limited space of the board having a rigid geometric shape. Especially noteworthy is the fact that each piece occupies and is allowed to occupy the same area of the board as any other piece. This indicates a certain degree of equality among the pieces.[12] I do not know whether this rule was introduced because of some special considerations rooted in the social structure or just to simplify the game.

---

[12] A desire to create equality among members of society arises not only from altruistic considerations but also from the fact that society with inequality seems harder to integrate.

Equality in chess is not limited to certain properties of the pieces. The basic structure of the game presupposes that both players are equal both in the quantity and the quality of the pieces they have as well their starting positions.

The fact that the initial conditions for both players are the same is evident in the present day version of the game - the shape of the pieces is the same for both players but the color which has no special symbolic meaning is different.[13] At the same time there exist some forms of chess where this feature is modified while all other rules are basically the same as in regular chess. Mongolian chess is interesting in this respect: it mimics the universal concepts of good and evil.[14] In keeping with their tradition the Mongols considered red (as well as yellow) to be masculine or spiritual color - the color of the sun; green was seen as feminine or earthy color - the color of the Earth. Consequently the base of the pieces belonging to the two players is colored differently: they are either all red or all green. (The upper parts of the pieces are painted with both colors perhaps to reduce the stress on the players. It may also signify that each player possesses characteristics associated with both colors). For a long time the Mongols considered Chinese to be their arch enemy. Their attitude is reflected in the shapes of the pieces. The expressions on the faces of the "kings" representing the Mongolian prince and the Chinese ruler are quite different. The queens differ in shape as well: one is in the form of a lama - a Buddhist lion and the other is in the form of a tiger-devil borrowed from the primitive pre-lamaic religion. The shapes of the camels, horses, and carriages are the same for the two players. But the pawns are very different: they are carved into shapes of different birds meant to express the conflict between the spiritual and the material, between the heavenly and the earthy.

Although the initial position and the number of pieces is the same for both players the game does reveal differences in their innate abilities and knowledge. Equality in the starting position and number of pieces which the two players have at their disposal vanishes as the game progresses and eventually leads to a victory (defeat) for one of the sides. Equal starting conditions are reinstated at the start of each new game.

---

[13] A book by Mackett-Beeson, A.E.J., Chessmen. New York: G.P.Putnams Sons, 1968 shows photographs (some of them in color) of chess sets used in different countries. These photographs are accompanied by a brief note on the history of some of the forms of chess, the materials from which the pieces were made, and so on.

[14] Camman, S., "Chess with Mongolian Lamas", Natural History, November 1946, pp. 407-411.

Digression 1.

The concepts of equality and inequality in chess stimulated me to apply these general ideas to social systems. The problem of social inequality is rooted primarily in the fact that initial material resources are not distributed equally among the members of society. All other conditions being equal people who start out with more material wealth can achieve more in life than people who started out with less. Here inequality in wealth should not be confused with inequality in individual abilities - inequality in drive, energy, ability to perform various tasks, and intelligence - a quality which integrates all these characteristics among themselves and the environment. The relative impact of innate abilities, i.e. the genetic code and the time spent in the womb as opposed to the role of the environment is a very complicated issue. This age-old debate is outside the scope of the present work. Personally, I think that innate abilities are crucial in the formation of an individual and that they should be taken into account in our attempts to resolve social problems.

Finally, inequality in material wealth and individual abilities should not be confused with the fruits of one's labor which results in inequality in current income. This kind of inequality may be caused by the two-fold inequality mentioned above, i.e. inequality in the initial conditions and in innate abilities.

In his wonderful work called "white revolution", V. Z. Jabotinsky[15] examined all these aspects of social inequality from a very interesting perspective. The "socialist" approach to the problem of inequality is limited because it tries to make both the initial conditions and the current income of all individuals equal once and for all. The drawbacks of the "bourgeois" method is that it tends to perpetuate these inequalities.

Jabotinsky thinks that the most effective solution is to combine the socialist principle of equality in the starting conditions with the bourgeois principle of inequality in current income. He noted that equality of initial material resources is important to ensure a favorable start for every individual. At the same time he emphasized the need for inequality in the distribution of wealth in order to stimulate people to develop their innate abilities. How can we combine these two principles? Jabotinsky showed that certain laws on the redistribution of wealth will help society move in the direction of equality in the initial conditions without resorting to bloody revolutions; here the spirit of achievement will be fueled by the inequality in incomes.

---

[15] Jabotinsky, Z., Selected Works. Jerusalem: Alila library, 1978.

To illustrate his idea Jabotinsky borrows an example from the Old Testament. There is a law in the Old Testament which states that the land purchased at some point in time must be returned to its owner after a period of 50 years. This is where the Hebrew word "jubilee" comes from. It found its way into the Roman lexicon and is now part of many European languages.

This kind of a "white revolution" is different from the "red revolution." The latter involves violent means used in order to change the initial conditions of the people.

Jabotinsky notes that the laws of the Old Testament were appropriate only under the then prevalent historical conditions but the principle used is still of interest. It can be applied to new situations by enacting appropriate laws which would ensure a peaceful solution to the problems arising in the course of development.

§3. The rules and the operator of the game.

Now, let us come back to chess. I want to talk about the process aspect of chess. As far as the set up of the game is concerned it manifests itself in the rules of the game. Rules enable players to preserve the unity of their noble goal: to enjoy the game while keeping the spirit of competition (in terms of the means) alive. These rules which are rather extensive pertain primarily to the moral requirements placed upon the players. Oftentimes they are not taken into account because of the emphasis on the rigidity of the rules of the game itself.[16]

The use of chess as an allegory of moral behavior has a long history. A book devoted exclusively to this topic appeared as early as14-th century. It was written in Latin by Jacobus de Cellosis and was called Libellus de Moribus Hominum et Officiis Nobilium. It was translated into all the European languages 50 years after its first publication.[17]

So, what are the rules of chess that are so important from the moral point of view?

First of all, the game is governed by a set of strict rules which both players must agree to obey. In chess battles the role of referee-judge is rather minute compared to such sports as football, basketball, boxing, etc. The gap between chess and real life is evident form the fact that in many cases there are no rules guiding the behavior of the conflicting parties. Even if such rules do exist there may be no referee to spot the violations.[18]

Rules of chess presuppose that the outcome of the game depends solely on the strength of the players. His majesty chance plays no explicit (or as such) role in the game. This makes chess different from many other games - card games, dominoes, lotto, roulette, dice, backgammon, etc. The role of chance in these games is big enough to determine the outcome of at least one round. These games require a large number of repetitions in order to eliminate the the luck factor.[19] To a certain extent these considerations also apply to chess as we know it. White pieces received by chance give the player certain first move advantage in that particular game. The "first move" advantage is largely neutralized when players play many games with each other.

On occasion the first move privilege does become significant even in a long match between the same players. World championship matches have demonstrated this a valid concern in cases where the last game turned out to be decisive.

Jumping ahead I would like to note that the importance of the first move stems in part from the fact that the starting position has many pieces constricted in movement. First move advantage could be neutralized by introducing a rule called a "tabia". This rule would have the game start at a particular position judged to be equal by both players.

Present-day rules of chess which eliminate the role of chance in the course of the game itself were not always the same. At one time a die was tossed before a player made a move. It seems that during the middle ages when Europe made an attempt to eradicate gambling the role of chance as such disappeared from the game.

---

[16] In his book on Stalin Trotsky gives the following account of the moral relationship between the player and the game:
"The laws of political mechanics once formulated by Machiavelli were for a long time considered extremely cynical. Machiavelli viewed the struggle for a power as a chess theorem. For him, moral dilemmas do not exist just as they do not exist for a chess player or for an accountant whose task is to do the most expedient thing under the circumstances." Trotsky, L., My Life, 1985, p. 14.
I disagree with Trotsky when he says that chess players face no moral dilemmas: the morality of a chess player should not be reduced to the way he moves the pieces. Psychology plays an important róle in one's play and sometimes one oversteps moral boundaries. If we compare the behavior of B. Spassky towards B. Fisher in their world title match in Iceland with the behavior of A. Karpov towards V. Korchnoi and G. Kasparov in their title matches, we see the impact of one's moral stature on his conduct. Morality is also important in the course of the game itself: for instance, whether one player "terrorizes" another with his rude behavior which may be legal but is morally unclean thus ruining the aesthetic side of the game.

[17] Michigan University Library has a microfilm of the work by Fuller, C., A Critical Edition of Le Jeu Des Eschs, Moralize. Ann Arbor: University Microfilms International, 1976, which has a detailed analysis of the French translation of this book.

[18] Huizinga,J., Homo Ludens. New York: Roy, 1950.

[19] So it is not an accident that international competitions only allow games where the role of chance (when it can favor one participant more than another) as "an additional player" is reduced to a minimum. Poker tournaments are permitted since each round takes a short time to complete so a lot of rounds can be played within a reasonable period of time. Statistically this helps eliminate the luck factor. Bridge competitions are organized in a very "tricky" way: each side has two teams which play simultaneously: in each round, one team plays with its own cards and the other has the cards of the first team's opponents.

We should mention another important provision having to do with the chance factor: at each moment of time each side knows the position of the opponent's pieces and therefore every move made by the opponent up to that point. This kind of information can be helpful in predicting the future moves of the opponent.

One form of chess called Kriegspiel lacks this rule. Kriegspiel is the same as regular chess except neither player knows the moves made by his opponent. Therefore, each player must be told by a referee who keeps track of the game whether or not his moves are legal.

Absence of chance per se and complete knowledge of the opponent's position idealize chess as compared with real world conditions where these factors may play a significant role.

Let us now examine some rules of chess as far as they reflect society's moral code. First of all, there is a rule which states that a player is not allowed to remove his own pieces from the board, i.e. to "kill" the piece "himself." One can offer a piece as a sacrifice leaving it up to the opponent to decide whether to accept or reject the offer. This rule is very strict: a player is not allowed to capture his own piece even when its removal from the board would prevent an immediate defeat or lead to an immediate victory. This rule is somewhat reminiscent of the moral principle which states that those in power cannot kill their own countrymen under any circumstances. At the same time homicide is permitted in military combat - both those killed by the enemy directly and those offered to the enemy for the ruler's own benefit. This moral principle is quite realistic. It is fully applicable to such countries as Sweden, Denmark, Switzerland, and others. These countries not only forbid the murder of innocent people for whatever reasons but they have also done away with capital punishment for criminals; at the same time these countries have armies and are ready to kill their enemy and sacrifice their own population.

Another rule of chess states forbids the use of any piece captured from the opponent: the piece leaves the board once and for all. This rule reflects a belief that each participant in the combat should remain loyal to his country and even if captured he should not be fighting against his own army on the side of the enemy. But a Japanese version of chess called Shogi[20] permits the use of the captured piece as your own (it is somewhat inferior because it is not allowed "to advance in rank" immediately; more on this below). The

---

[20] Games of the World, ed. by F. Grunfeld. New York: Ballantine Books, 1975, pp. 77-83.

captured piece can be placed on any unoccupied square on the board.[21] Perhaps this rule reflects the conditions prevalent in the wars which took place between the Japanese feudals during the period when Japan was a less centralized country.

Stalemate rule is rather curious. The winning side must conform to another constraint of avoiding immobility of the opponent's king if all his other pieces are also immobile. The roots of this rule are unknown to me. Perhaps it arose in the middle ages if we interpret the rule as preventing situations where an attacked knight has no place to move before being directly struck and lacks mobile troops to continue the struggle. In other words "Do not hit a man when he is down".

Finally I want to note that the rules of chess restrict the capacity of a piece to change its "social status," i.e. its role in the game. The attributes of a given piece are fixed once and for all. It may gain or lose in strength depending on its position on the board or the positions of other pieces. (For instance, a bishop can only make diagonal moves. On an empty board, it can make anywhere from 7 moves if placed on the edge of the board to 13 moves if placed in the center of the board).

A pawn represents an exception to this rule. It can "be promoted." In fact once it reaches the opposite edge of the board it must be promoted to any other piece including a queen but not a king.

We should again note that Shogi provides far greater opportunities for the pieces to change their "social status". Pieces can be promoted once they reach one of the rows near the edge of the board. They also do not have to change immediately upon reaching a certain row: a player can change them any time after they have acquired the right to "advance."

Everything said so far is related to the social and moral attitudes toward an individual. What is the role of innate abilities in determining a person's life? Does social mobility depend on personal success?

Let us now examine the notion of an executor-operator of the chess game: is this a man, a machine, or some combination of the two? Chess presents us with all of these combinations: man against man, man against machine, machine against machine. Here, the term machine stands for any artificial device which plays the game. History of chess is abundant with examples of these kinds of machines turning out to be "imposters" - in one case a man was put inside a machine but it looked like the moves were made by a mechanical device.

---

[21] This rule as well as the "advancement in rank" rule are easy to implement in Shogi since all the pieces have the shape of a pentagon. Pieces are distinguished by the hieroglyph drawn on top of them. The direction in which the pentagon shaped piece points determines its "owner."

So, what are the requirements placed upon the operator of the game considering that each game must be played in real time, i.e. only a finite amount of time is allotted to each game?

Jumping ahead a little I want to note that a machine algorithm for playing chess does not have to imitate that of a man if only because a machine is capable of calculating variations much faster than a human being, it does not get "tired," it can house a huge memory, and so on. At the same time a number of very difficult problems attributable to the role of individual experience arise when we try to design a machine algorithm of the game. There are examples of technological devices whose principles of operation differ from the analogous objects occurring in nature.[22]

The other approach assumes that computer chess programs must be based entirely on the principles of the game elaborated by man. Much can be said in support of this approach. For instance, there is an entire area of science called bionics which deals with design of new technologies based on the principles found in the living systems (the principle used by the bats to navigate was used in the development of radar); late 80-ies witnessed the development of the idea of designing computers capable of pattern recognition which would mimic the neuronetwork of the brain as opposed to fast and powerful sequential computers.

It is hard to say at this point which approach or which of their combinations is most beneficial to the development of a chess algorithm. In any case experience with computer chess has revealed the need to combine the ideas of incorporating both the human experience and the special capabilities of a machine.[23]

At the present time, research of chess algorithms is linked primarily with computers; an elaboration of such an algorithm for the human players is not considered to be a crucial issue.

---

[22] Stevenson's genius consists of having discovered a new principle for transforming the back and forth motion of the steam engine into the rotary motion of the wheel. An attempt to build a steam locomotive in the image of a horse - steam engine being its heart, ended in a failure.

[23] Adelson-Velskii, G., et all, Machine Plays the Game, Moscow: Nauka, 1983, pp. 60-71.

Important results have been achieved in the field of computer chess. The best chess programs of today play almost as well as grandmasters: on September 25, 1988 chess program Hitech won a match with grandmaster A. Denker by a score of 3 1/2 to 1/2.[24]

Sceptics who did not believe in the future of computer chess such as the philosopher Herbert Dreyfus from Berkley who claimed that a computer will never beat even a ten year old have been put to shame.[25] But overly optimistic predictions have not been fulfilled either. For instance, in 1957 Professor H. Simon from Carnegie-Melon University made a statement that in 10 years computers will play better than the world champion. But, alas!

Still, I tend to side with the optimists. Although their expectations may have been too high they stimulated the development of computer chess and achieved a number of great victories on the way. Therefore I feel supportive of the statement made in 1986 by Hans Berliner from Carnegie-Melon that by the year 1990 his chess program has a 50% chance of beating a world champion, both among people and machines. A pledge of his success lies in the fact that he created the first computer program to beat the world champion in a board game - his backgammon program. Also, by mid 1986 his chess program Hitech was considered the ninth best player in Pennsylvania and 200-th best player in the United States and it still has a lot of room for improvement.

§4. The genesis of the game.

Now let us move on to the historical aspect of the game. Chess is one of the most ancient games. Although traces of games similar to chess have been found on ancient Egyptian sculpture[26] most scholars agree that the present day version of chess (or similar version to be precise) first appeared in India in 7-th century A. D.[27]

Some attribute the invention of chess in India to one of the game's aforementioned functions, namely it simulates military combat. It is held that the Buddhists who were quite influential in India at that time were able to introduce the game to the people. It was in harmony with their ideology which renounced violence and saw chess as a substitute for war.[28] There were even instances when two European countries which were at war with each other settled their disputes at the chess board. Unfortunately the reverse also took place. A player, seeing inevitability of unfavorable outcome would take some drastic measures against his opponent, even murder.[29] As a matter of fact during the middle ages chess pieces could have been used directly as a weapon. Considering their weight and the size of the board a "mutual exchange" of material could lead to some very unsettling results.[30]

---

[24] Schonberg, A., "For First Time, a Chess Computer Outwits a Grandmaster in Tournament", The New York Times, Sept. 26, 1988.

[25] See Time, Oct. 28, 1985, p. 88.

[26] Lasker, E., Lasker's Chess Manual, Moscow: OGIZ, Fizkultura i Turizm, 1937, p. 5.

[27] Games of the World, ed. by F. Grunfeld. New York: Ballantine Books, 1975, p.63.

[28] Koestler,A., The Hill of Achilles. London: Hutchinson of London, 1974, p.198.

The present day form of chess is special in its variety of material objects - pieces, their ability to move around, and rules on how they ought to interrelate. This distinguishes chess from many other games that are much more limited in their potential. At the same time, chess in its present-day form is a rather conservative game compared with other systems (especially social ones) in which new objects and new rules continue to appear. Not only is the variety of pieces limited to what it is now but the current rules of the game prevent new pieces or new rules from being introduced. If such changes do take place they are very rare and usually lead to the creation of a new version of the game.

History of chess is rather rich in this respect since the game has undergone some profound changes in the course of its evolution. Here we should note that there are several major forms of chess now in existence. Beside the game properly known as chess there is Chinese chess and Japanese chess known as Shogi.

At the same time, chess players are trying to invent new original forms of the game.[31] Along with modifying the rules of the game or inventing new kinds of pieces with unusual abilities new shapes for the board have also been proposed. The board can simply be increased or decreased in size (maxichess and minichess), its shape can be changed keeping it two- dimensional (like hexagonal chess), or it can be constructed in three dimensions (cylindrical, spherical, or torus chess); theoretically, we can even think of multi-dimensional chess. These new forms of chess were not merely a fruit of idle imagination. Many of these new forms of the game (with different boards, new rules and new pieces) were actually implemented and some were even played at a competitive level.[32]

---

[29] Fuller, C., A Critical Edition of Le Jeu Des Eschs. Moralize. Ann Arbor: University Microfilms International, 1976, p.15.

[30] Koestler,A., The Hill of Achilles. London: Hutchinson of London, 1974, p.199.

[31] Gik, E. J., Chess and Mathematics, Moscow: Nauka, 1983. Examples of modifications given here are taken from the two chapters of the book - "Games played on unusual boards" and "Magical chess".

[32] Capablanca who invented maxichess played a match with a Hungarian grandmaster Marotzy. In 1929, the first tournament in hexagonal chess was held in Europe.

In summarizing the above discussion concerning the formulation of the problem (the role of chess in the solution of the problem shall be discussed below) chess represents a model which not only simulates reality, especially its social aspects, but also has a big constructive impact on the development of the world.

Chess is an example of competition which is subject to a rather elaborate set of rules: the board, the pieces, and the rules of their "conduct." Chess also represents an interesting model of a social system where, on the one hand, the initial conditions of the two players are the same and they can be reconstructed at the start of each new game and on the other, the outcome of the game depends solely on the individual abilities of the participants.

Nevertheless chess presents a rather limited view of the world because it has less variety that real systems: "participants" (pieces) and rules for manipulating them are given, i.e. neither the pieces nor the rules can be changed within the system (not to be confused with new possibilities introduced from the outside as novel versions of the game). Also the game possesses a final and a rather limited goal (to capture the king without any concern for the other pieces); the rules are reduced to a form of competition where chance plays no explicit role and the positions of all the pieces are known to each of the participants at each moment of time. A ban on removing one's own piece (taking them off the board) under any circumstances, pieces being limited in their ability to change their "social status," and so on are all rules which make chess further removed from reality.

Discussion which follows focuses on the solution of the problem of chess. We shall examine a number of general ideas which come up despite or rather as a result of the simplicity of the game.

Chapter 2. The game at the global level.

§1. Optimization vs. chance. An inexact problem.

So the game of chess as a problem has been formulated. Its definition is exhaustive: the goal, the initial conditions and the rules governing the operations allowed in the course of the game.

We are presently faced with a task of designing and implementing a method or an algorithm for solving this problem. Design of a chess algorithm requires a clear understanding of the nature of the operator or the executor of this algorithm since this algorithm must be executed in real time. Our subsequent discussion is based primarily on the recent developments in computer chess and, where pertinent, the experience accumulated by the chess players over the course of many years.

Design of a chess algorithm involves furnishing a given "player" with an optimal "trajectory" of development, i.e. a sequence of moves which maximizes "his" objective function. Formally this algorithm must maximize a function whose arguments can assume the values 1, 1/2, or 0 - a win, a draw, and a loss respectively. Theoretically a search for an optimal strategy for one player must be accompanied by the same kind of search for the other player since it is assumed that the opponent is also rational and plays the "best" possible moves. Finding an optimal strategy in chess which is a zero sum game means designing a max-min algorithm where a maximum gain for one player is equal to a minimum loss for the other.

Now, before attempting to implement this kind of an algorithm let us ask ourselves the following question based on one of the principles of "non-constructive mathematics": "Does an algorithm of the game exist at all?"

Theoretically, such an algorithm does exist since the number of mating positions attainable during the course of the game is finite. True, the number of such positions is huge. Even if we assume that the game only lasts 40 moves (5-6 thousand moves represents the limit if positions are not allowed to repeat) there are 10^{120} possible variations from the starting position. It would take a computer 10^{90} years to go through all these variations even if it spends one micro-micro second on each one.[33] So although an exhaustive search would require only a finite amount of time it is irrelevant to the human time scale.

---

[33] Shannon,G., 1950, "Programming a Computer for Playing Chess", The Philosophical Magazine, vol.XLI, 1950, pp. 259-260.

Therefore a real solution to the problem calls for an ordered search with the number of operations bounded by some reasonable number. Ideally an ordered search ensures an optimal outcome of the game. This certainty stems from the optimization principle which consists of finding an optimum point by means of a completely ordered procedure, i.e. a procedure proved to be convergent and which searches through all the variations eventually eliminating all the inferior ones. In the last few decades research in the areas of linear (including integer solutions) and dynamic programming has proved the existence of algorithms which lead to an optimal solution in a finite number of steps. There are rigorous proofs of the fact that these procedures converge to an optimum point.[34]

Chess, it turns out, is too complex for present day optimization algorithms. Unfortunately chess literature that I am familiar with does not discuss the specific reasons why chess does not yield (in principle) to these optimization algorithms.

We should add that an effective algorithm of the game cannot be reduced to elaborating an optimization algorithm. A useful tool in designing an effective algorithm based on selective search procedures takes into account the specific structure of the board and the rules of the game. K. Thompson, for instance, designed an algorithm for 5 piece (including kings) endings which generates an exact move corresponding to the best possible course of action.[35] These retrograde algorithms (i.e. they start from the end - from winning positions) are basically based upon an exhaustive search and so do not, strictly speaking, belong to the class of optimization procedures. Nevertheless, the search is greatly reduced by taking into account the symmetry of the board. Under certain conditions a solution corresponding to a particular square can be extended to three other squares.

We should note in passing that algorithm efficiency can be further improved through the use of specialized chess computers.

My subsequent analysis deals only with those chess algorithms which employ all kinds of methods of truncating the search tree and which do not rely directly on board structure, rules of the game, or the speed of the computer.

Since an optimization-like search of all the variations is denied making random moves may seem like the only other way to play. Indeed we rarely know all the consequences of a given move so we should allow his majesty chance to decide the move for us. Since the number of moves is finite and present day rules limit the game to approximately 6000 moves before it exhausts itself completely, random move generation would lead to an end of the game within a reasonable period of time. But this method deprives the game of all its meaning. The game loses its cohesive structure.[36]

---

[34] It can be proved that algorithms in linear programming must perform an exhaustive search in order to find an optimal solution to a certain set of problems. In practice such a search is not required for an optimal solution.

[35] Thompson, K. "Retrograde Analysis of Certain Endgames", ICCA Journal, Vol. 9, No. 3, 1986, pp. 131-139.

I do not mean to play down the role of chance in chess completely. Research of algorithms of mathematical programming (with proved convergency) has shown that under certain conditions we can achieve good results by combining ordered and rigid procedures for determining the best course of action at the start of the iterations (such as the simplex method) with a random method (like the Monte-Carlo method) closer to the end of the program.

We thus face the problem of designing an approximate algorithm of the game hoping that it would produce good results within a reasonable period of time. Unlike optimization or exact algorithms, these algorithms could be called good-non-optimal or approximate-inexact algorithms depending on the attitude of the scholar. Von Neumann called these indeterminate, heuristic methods of play "good chess."[37]

M.Botvinnik[38] referred to the class of problems where such algorithms are used as inexact problems. Subsequently Botvinnik gave a more precise definition of the inexact problem. I quote it below.

> "We now say that an enumerative task is inexact if it solves a problem by minimax methods on a truncated search tree (see Glossary of Terms). The concepts of minimax procedure and search tree are well known; the concept of the truncated search tree may need explanation. Problems soluble by the formation of a tree of all possibilities (elementary actions) may differ in difficulty and may give rise to search trees of different sizes-small, large, or even infinitely large. If the resources of our information-processing device (speed and memory) are so limited that we cannot form the tree and search it exhaustively, we must either abandon the task or be content with an inexact (i.e., approximate) solution. If an inexact solution is acceptable, we limit the depth of the variations; the use of a depth-truncated tree and the acceptance of an approximate solution make the enumerative problem inexact. The definition of an inexact problem is therefore inextricably bound up with the general method of solution and with the resources of the information processing system being used. If we can apply the minimax procedure to the complete tree, the problem is exact. (Problems may of course be solved exactly by other methods, e.g, by the use of equations or exact algorithms, and they may be inexact under other definitions, as when approximate solutions to equations are used."[39]

§ 2. Algorithms of inexact problem - general description.

Generally speaking, final product, structure, function, process, and history, all being parts of a chess algorithm, must fulfill the following requirements:

Functions. a) to induce the opponent (for instance, influence opponent's choice between an aggressive and a defensive strategy).

b) to ensure at each stage of the game the capacity of one's position to adjust favorably to unforeseen moves.

c) to ensure at each stage of the game the capacity of one's position to absorb the impact of unforeseen moves if they happen to be disadvantageous to the player in question.[40]

---

[36] To get pleasure from the game itself, the search must be ordered in some creative fashion. The fact that computers receive no such pleasure is a result of them lacking a program of creation, i.e. they do not form these programs of creation by themselves. It seems to me that an emotional mechanism is not a prerogative of man or animals; it could be implemented in machines provided they posses an elaborate mechanism of creation.

[37] Von Neumann, J. and Morgenstern, O., Theory of Games and Economic Behaviour, Princeton, 1944, p. 125.

[38] Botvinnik, M.M., Algorithm of a Chess Game, Moscow: Nauka.1968.

[39] Botvinnik, M.M., Computers in Chess: Solving Inexact Search Problems, Springer-Verlag: New York, 1984, p.1.

[40] Generally speaking, this approach to algorithm design differs from the adaptive approach which is based primarily on responding to changes in the environment by changing the internal organization of the mechanism of adaptation; here, only one of the functions mentioned above, namely the third one, is fulfilled. R. Ackoff suggests a new broader approach to the problem of adaption which is basically similar to my own. This is what R. Ackoff writes on this topic:

"To adapt is to respond to an internal or external change in such a way as to maintain or improve performance. The change to which adaptation is a response may present either a threat or an opportunity. For example, the appearance of a new competitor may present a threat; the disappearance of an old one, an opportunity. Both require an ability to detect changes that can or do affect performance and to respond to them with corrective or exploitative action. Such action may consist of a change in either the system itself or its environment. For example, if it suddenly turn cold, one can either put on additional clothing (change oneself) or turn up the heat (change the environment). Furthermore, the change to which adaptation is a response may occur either by choice or without it. The demise of a competitor, for example, may occur independently or because of what a corporation does.

The concept of adaptation used here is much richer than the one used in association with the theory of evolution. In that theory, adaptation refers to only involuntary responses to external changes, and the responses consist of internal changes. This restricted connotation of the concept derives from the fact that the theory of evolution is preoccupied with nonpurposeful systems, and when it deals with purposeful systems it is not concerned with their purposefulness. Here we are preoccupied entirely with purposeful systems and their purposefulness". Ackoff, R., Creating the Corporate Future. New York: Wiley, 1985, pp. 170-171.

It seems to me that Ackoff's ideas on adaptation and their connection with the functions of an algorithm outlined above can be elucidated if presented in a more precise matrix-like.

Final product. In order to fulfill the above requirements we must know how to evaluate the final product generated at each step of the game. First, we shall describe the structure of the final product. The most difficult questions pertaining to inexact algorithms arise at this point. Future analysis shall focus primarily on the methods used in generating the step's final product.

Structure. The structure of the algorithm involves three dimensions: it has "length", "width", and "depth."

"Length" of an algorithm refers to its division into stages, stages into substages, and so on until we finally reach some primary step. The primary step (for instance in Hitech chess program) could represent a move chosen by each square of the board and indicating the best course of action leading to a potential placement of a piece on that square.

"Width" indicates the number of pieces included in the analysis of a given position. According to M. Botvinnik[41] the "width" factor can be split into three levels. (It seems to me that a fourth level can be introduced if the first level of Botvinnik's scheme is divided into two.) First level - identify the attacking piece. Second level - identify the stem pieces; these include the attacking piece along with all those pieces which, within the horizon of the attacking piece (same number of half-moves), help it achieve its goal. Third level involves the formation of a zone which M. Botvinnik defines in the following way:

> "Stem pieces do not act alone. Each is a crew of pieces of its own color who support it; there is also a hostile crew of pieces of the opposite color who hinder it. The ensemble of stem pieces and the two sets of commands of pieces of both colors from a field of play - and this is the second (intermediate) level of the control system ". (p. 34).

Fourth level deals with MM - mathematical mapping which includes several zones. It represents possible courses of action in the form of tree within which the actual search tree is constructed.

Finally, by "depth" I mean a hierarchy of programs. Besides the first-level programs, i.e. programs specifying what, how, from what, by which means, when and where,

Table 2. Methods of adaptation

Evaluation of the actual impact of the changes in the environment	Changes in the environment	
	Planned	Random
Positive		
Negative		

---

[41] Botvinnik, M.M., The Cybernetic Goal of a Game, Moscow: Sovetskoe Radio, 1979.

we must also have second-level programs which determine the first-level ones. There must also be higher level programs based on the experience derived from the game and the knowledge of the methods of solution of similar kinds of problems. The purpose of these programs would be to improve lower level programs including the first-level ones.

Process. Methods of implementing the process can be grouped into several classes; these classes can be split into orders, orders into families and so on.

Here, I just want to note that the reactive principle of move selection invokes player response to a given situation that is the same all the time: "input" determines "output". This principle avoids the need to search through all the possible variations. In other words, the reactive approach means that the initial state determines the choice of the function which is known from previous experience. This function, in turn, transforms a given state into a new state. A search principle of selection shall mean the following: first, a player formulates a problem of finding the best move and then looks for it by searching through different variations.

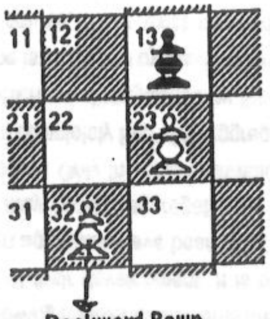
History: Information about past play, both your own and your opponent's is important in choosing one's strategy (aggressive - defensive), in predicting the most probable moves, in establishing the value of a position, etc.

I want to repeat that every algorithm must fulfill the above requirements no matter what its implementation of the function, the final product, the structure, the process, and the history are. Attempts to make these general rules more concrete and finding ways of implementing them under a variety of circumstances has resulted in a number of different algorithms of play. Solution of the problem can be twofold - implicit or explicit. Under an implicit approach, all requirements are part of past experience represented in the form of specific and independent recommendations. Under an explicit approach, these requirements are incorporated in the formulation of an individual problem explicitly. Implicit and explicit approaches to inexact games are based on the reactive or the search principles of selection; naturally, these two principles can be combined.

§3. Reactive methods.

The typology of reactive methods introduced in the part of the book on large complex systems distinguishes between reflexes, instincts, images, and stereotypes. With regards to chess, a reflex is analogous to moving a piece away from a direct threat. An instinct corresponds to a situation where pieces are identified based on one single parameter, namely their ability to capture an opponent's piece. An image can designate such configurations as a backward pawn (see fig. 1) since it is bad on several counts: it can be captured, it creates a forpost for the opponents pieces which is protected by his pawn and which cannot be attacked by ones own pawns.

Fig.1



Backward Pawn

Stereotype refers to repeating the same move in a given position.

For my purposes, I shall distinguish only between a stereotype and three other reactive methods. Unlike a stereotype which encompasses a given situation in its entirely the other three methods take account only of some parts of a position. It seems to me that the widespread term "heuristic" is a good description of these three types of reactive methods.

Heuristics are valuable both for beginners as well as for advanced players. Naturally, heuristics' level of complexity varies depending on the level of play.

Having learned the rules of the game a beginner plays in a random like fashion. Having acquired a certain amount of experience, he begins to notice that some moves are good and some bad. Chess theory tries to formalize this experience and present it to the beginner in the form of recommendations. So the first steps in making a beginner's play more coherent consist of recommendations telling which moves are reasonable and even more importantly, which are not. These recommendations are similar in nature to heuristics. For instance, one of the first things learned by a beginner is not to place his pieces under attack or more precisely, not to blunder away his pieces; it is advisable to avoid making several consecutive moves with the same piece especially in the opening stage of the game, etc. Positive recommendations are also given; for instance, advance central pawns to gain control of the center and to develop the pieces.

The number of possible heuristics may grow with the level of skill. Some recommendations are valid only for the beginners. They are modified as far as more mature players are concerned. For instance, chess theory asserts that control of the center is not limited to its control by the pawns; hypermodernism introduced control of the center by opening up the flanks, i.e. by the bishops which assume extreme flank positions after a few preliminary moves. But hypermodernism requires considerably greater expertise on the part of the player.

I tried to point out a number of recommendations intended for beginners which contradict those given to the more advanced players in order to emphasize the contradictory nature of the heuristics themselves. We can assume that heuristics are given as independent statements. It is further assumed that these statements are always true. Some heuristics, such as "do not blunder" never violate these assumptions. But many heuristics are only true statistically which means that we can find a set of conditions under which they are not applicable. For example, there is a heuristic rule which disapproves of doubling one's pawns; yet there are cases where this rule does not apply (a player may want to use the open file for his heavy pieces). Moreover, heuristics may contradict each other. Therefore, they cannot be considered independently of each other in a given position.[42]

> "For instance, how can we make the heuristic "do not touch the pawns protecting the castled king" compatible with 'a very natural desire to drive away the opponent's trouble causing pieces or with the heuristic 'do not forget to make a window for your king" ?"[43]

Conflicts between heuristics are not unusual.

> "Many heuristics used by the players are actually incompatible with each other. Therefore, we must create the same heuristic procedures to determine the importance of each heuristic. We then choose some subset of them depending on the nature of the position in question. This particular trend in heuristic design has seen very limited development."[44]

---

[42] We have a similar situation with sayings and proverbs. Proverbs reflect wisdom and experience of many generations. But the unconditional nature of these sayings which sometimes results in their incompatibility makes it difficult to apply them in practical situations. For instance, a proverb "Two heads are better than one" has its opposite in the saying "Too many cooks in one kitchen". Two things must be done before we can apply these proverbs; we must either define conditions of their applicability or determine their priority in case their application to the same situation creates a conflict.

[43] Adelson-Velskii, G., et all, Machine Plays the Game, Moscow: Nauka, 1983, p. 70.

[44] Ibid.

Generally speaking, the reason for this situation with heuristics is the fact that chess theory as a whole is not based on one single concept but rather on a set of heuristics.[45]

From a methodological point of view, this means that recommendations-heuristics incorporated into an algorithm do not represent one single system: they are more like an assemblage (an aggregate) comprised of independent recommendations.

These independent recommendations comprising an aggregate are useful because of their statistical validity: they remain valuable even outside a given case. Each heuristic retains some "part" of its value even in situations where the heuristics are assigned priorities because their simultaneous use results in contradictions. The significance of the remaining "part" may become apparent in the future but at that particular point of the game its nature remains unclear. Therefore, situations which may give rise to contradictions among the heuristics should be avoided although in each specific case it may be possible to determine the priorities among the heuristics aposteriori depending on our goals and intentions. But this remaining "part" cannot be ignored in case contradictions among the heuristics are unresolvable and we have to resort to setting priorities among the them.

On the basis of general considerations we can say that aggregate structures comprised of a large and possibly growing number of elements are typical of many systems, especially in the beginning stages of their development. It is only after all the relations among the elements belonging to an aggregate have been uncovered that the latter turns into a system. Advanced systems may contain aggregates as well as subsystems (like systems, subsystems are also ordered). Only in certain extreme case and under certain conditions does an algorithm become completely free of all aggregates.

Now consider stereotypes which are distinguished by their holistic approach to the position. Effectiveness of a stereotype is confirmed by previous experience and occasionally by theoretical considerations. Of course, there exist no rigid deductive proof of the impossibility of replacing a given stereotype with a better one.

Chess players' experience with stereotypes involves getting acquainted with the games of other players and learning the lessons to be used in playing similar positions occurring in one's own games.

In principle, this approach seems very conducive to the development of chess algorithms. According to C. Shannon, this methods involves having a "dictionary" of all possible chess positions. There is a "right" move (somehow arrived at or suggested by a

---

[45] Ibid., p. 61.

grandmaster) associated with each position in the "dictionary." When it is machine's turn to move it just looks up the position and makes the appropriate move. The incredible number of possible positions - $64!/\,32!\,8!^2\,2!^6$ or approximately 10^{43} invalidates this approach.[46]

Obviously there are cases when algorithms incorporating previous experience (effective moves made in similar positions) are of value. The role of experience is especially prominent in grandmaster play and it can surely be implemented in machine algorithms.[47]

Still, it is hard to base one's play on remembering the best move associated with a given position because the number of possible positions is so large.

We should note in this connection that expert chess players are able to remember a rather large number of stereotypes since they memorize them in a very nontrivial way. Everything points to the fact that professional chess players memorize positions through their structure, i.e. groups of pieces making up certain configurations which H. Simon has called chunks. Psychological experiments conducted by A. de Groot were based on the hypothesis that chess player's memory is not purely mechanical. To test the hypothesis expert players and novices were given a position from a real game and allowed to look at it for five seconds. Some positions were quite complex with 20-24 pieces on the board. Professional players, master level and above were much superior to amateurs in this category. In random positions professionals and amateurs exhibited approximately the same abilities.[48]

It seems to me that expert players do not merely memorize a position but link it with the logic of possible continuations[49]. This leads me to conclude that an effective use of stereotypes calls for positional analysis characteristic of selective methods aimed at searching out novel solutions.

Moreover even if memory of the past is retained there is no guarantee that a given move is really the best one: it may be good; meanwhile best is the main enemy of the good.

---

[46] Shannon, G.,1950, "Programming a Computer for Playing Chess", The Philosophical Magazine, vol.XLI, 1950, p. 260.

[47] See Chess Skill in Man and Machine, ed. by P.Frey,1977, New York: Springer-Verlag.

[48] de Groot, A., "Perception and Memory vs. Thought: Some Old Ideas and Recent Findings", Problem Solving, ed. by B. Kleinmuntz. New York: Wiley, 1966, pp. 19-50.

[49] This reminds me of poets who remember a lot of poetry thanks to their understanding of the laws of poetic construction and remembrance of original key words which allows them to reconstruct the flow of a poem. Poetic process of reconstruction may proceed very fast and remain unnoticeable to an outside observer. In other areas this process is far slower. For instance, in mathematics one can memorize a formula or the logic of reasoning behind a whole class of formulas. The latter method may require substantial amount of time in order to come up with a specific formula.

We would therefore like to be able to group different positions into classes, especially by using chunks. Differences among positions belonging to the same class would be so minor that a move which proved to be good in one of them in the past can be played in all the positions belonging to that class. But it is rather difficult to define all the criteria to be used in deciding whether or not a particular position belongs to a given class. These problems are well known in psychology in the theory of gestalts and in artificial intelligence as it pertains to pattern recognition. Here selective methods come to the rescue. These methods are typically aimed at searching out new continuations from a given position. This approach requires a holistic evaluation of a position since there must be a criterion to judge which of the possible positions is best. Note that different positions being close in value does not mean that they call for the same methods of play. It merely indicates that they have equal predisposition towards future development.

§ 4. Search solutions.

We shall start by examining various types of optimization algorithms. This will be beneficial in subsequent application of these algorithms to inexact problems as well as in finding their most pragmatically feasible subclasses. In some extreme cases when parts of an inexact algorithm can be reduced to an exact problem these methods can be applied in their entirety.

Theoretically there are three different methods of constructing an optimization algorithm. The first method involves representing the entire game as a system of equations (inequalities) and then finding the values of the variables which correspond to the optimal value of the objective function. But representing a chess game as one single system of equations (inequalities) linking its initial state with the final goal seems an impossible task.

The second method of solving optimization problems is to

> "... to replace the actual problem space with a very much smaller space that approximates the actual one in some appropriate sense, and then apply the classical theorem to the smaller approximate space."[50]

This means that a problem like chess is to be broken down into a series of elementary parts with each part assigned a local problem whose solution will ensure a global optimum. The major idea behind this approach is the following: on the one hand, each local problem is given some degrees of freedom (choice) within the framework of the respective methods of transforming inputs into outputs. On the other hand, each local problem cannot be linked directly with the final, global goal associated with the problem as a whole (except for the final local problem); also the constraints of each local problem (except the first one) derived from the solution of a preceding local problem are not known. Therefore the formulation of a local problem requires some indirect information to be used in defining its local criterion of optimality. This criterion must guide the process towards a global optimum for the system as a whole. The criterion must also be implemented in a condensed form so that the local problem does not become too cumbersome and unsolvable for a given operator.

---

[50] Simon,H., "Theories of Bounded Rationality", Models of Bounded Rationality, vol.2. Cambridge: The MIT Press, 1982, p. 413.

Methods for solving these types of problems are known as algorithms of dynamic programming and decomposition principles of dynamic programming. These algorithms initiate search for the optimum point at the end of the game and gradually, using recursive procedures work their way to the the initial state. This approach ensures that the arrived at local solution is definitely an optimal one as far as the "future" is concerned since the procedure is structured "backwards" from the "future" to the "present." Nevertheless, methods of dynamic programming have limited application being of practical use only for problems having a small number of variables.

Other search algorithms for solving inexact problems are similar to the methods of dynamic programming: a gradual, step by step solution of a local problem coupled with a criterion for evaluating the "strength" of a given position and the constraints generated by the solution of the problem at the previous iteration. One major difference between these search algorithms and dynamic programming is the direction of motion which is, in the former case, from the "present" to the "future."

But this method does not take the game to completion or even to some intermediate stage - the horizon - which ensures success for this would, in the opinion of some chess players require looking 15 moves ahead especially in the middle game. This is the source of uncertainty as far as solution of a local problem leading to the desired outcome if one starts at the "beginning".

Especially interesting in this respect is the mathematical concept elaborated by Atkin.[51] Atkin's concept hypothesized about solve dynamic problems using algorithms which proceed from "beginning to end". I know of no attempts to apply this concept to chess.

---

[51] Atkin, R., 1972, "Multi-dimensional Structure in the Game of Chess", Int. J. Man-Machine Studies, 4, pp. 341-362;

Finally, I want to examine the so called decomposition methods of optimization. Basically, what makes them different from dynamic programming is that each local problem can be decomposed into sublocal problems which can be broken down further. Hierarchical organization of problems allows for completely new methods to be introduced: solution of local problems belonging to the same level can proceed in a parallel fashion while local problems belonging to different levels are related to each other in a sequential manner.

The idea of organizing local problems in a hierarchical fashion is used rather extensively in inexact chess algorithms. But here, unlike decomposition methods of optimization, hierarchy does not provide for parallelism. Rather it is used to set up some kind of an intermediate goal within a sequential process. This helps coordinate sublocal problems which are also structured sequentially. Hierarchy is also used in breaking up a sequential process of play in order to classify some aspects of the game (more on this below when we analyze different stages of the game). Judging by the brief description of chess program Hitech[52] its algorithm incorporates the idea of decomposition but it is limited to one move. At every move each square that is potentially a landing ground for some piece suggests to the "coordinator" what it thinks is the best alternative from among the ones leading to the placement of a piece on that square. The coordinator in turn hands down its evaluations of actions to each individual cell. The process is executed in a parallel fashion.

Our subsequent discussion of search methods used in the development of "inexact" chess algorithms is based on well-known chess programs having a rigid sequential organization of the local problems. Even if these algorithms do utilize a hierarchical principle it is supplemented by a sequential analysis of the problem.

Taking into account one single aspect of the problem, namely the multi-stage nature of the process, leads to the following question: "Should the local problem be uniform throughout (for all the moves it is supposed to handle) or should it vary?" Experience has shown that the set up of a local problem which is being solved as part of a global problem can vary. It is beneficial to vary the structure of a local problem since this makes for more effective solutions. At the same time, the solution of the global problem becomes more complicated. A heterogeneous algorithm, unlike a homogeneous or a unified algorithm which handles only one type of local problems, needs additional information from the local

---

Atkin, R., 1975, "Multi-dimensional Approach to Positional Chess", Int. J. Man-Machine Studies, 7, pp. 727-750.

[52] See Berliner, H., and Ebeling, C., "The SUPREM Architecture: A New Intelligent Paradigm", Artificial Intelligence, 28, 1986, pp. 3-8.

problems to ensure their subsequent integration into a single whole.

No matter what the set up of the local problems is we can never avoid the general question of the value of a given position in terms of its impact on the future of the game.[53] In fact, it will become clear in our subsequent analysis that evaluation of the current position is crucial in setting up local goals even in combinational style of play.

Analysis of a state, i.e. current position in the game, is vital. Such analysis should aim at establishing the proper set of parameters and their respective weights which should enable us to express the integrability of a given position with the game as a whole.

Moreover, there exists a trade-off between depth of search and complicated weight function.

> "Extrapolating from several microcomputer implementations at moderate speed with correspondingly constraint depths of search and complexity of evaluation function, the author is lead to conjecture that there is no difference - between deep search and superficial evaluation, and shallow search and complicated evaluation, though deeper search may provide stronger resistance to human opponents."[54]

These considerations form the foundation of our analysis of the structure of a position. Let parameters determining the value of a position be called parameters of integration. The next chapter examines these parameters in greater detail.

---

[53] "The problem confronting a chess player whose turn it is to move can be interpreted in either of two ways. First, it can be interpreted as a problem of finding a good (or the best) strategy - where "strategy" means a conditional sequence of moves, defining what move will be made at each successive stage after each possible response of the opponent.
Second, the problem can be interpreted as one of finding a set of accurate evaluations for the alternative moves immediately before the player.
From a classical standpoint, these two problems are not distinguishable. If the player has unlimited computational power, it does not matter whether he selects a complete strategy for his future behavior in the game, or selects each of his moves, one at a time, when it is his turn to play. For the way in which he goes about evaluating the next move is by constructing alternative complete strategies for the entire future play of game, and selecting the one that promises the best return." Simon,H., "Theories of Bounded Rationality", Models of Bounded Rationality, vol.2. Cambridge: The MIT Press, 1982, p. 412.

[54] White, J. F., "Querg Chess", ICCA Journal, Vol. 11, No. 2/3, 1988, p.72.

Chapter 3. Evaluation of a state: parameters of integration.

§1. Analytical representation of the holistic effect.

Evaluation of the current state must reflect the state's overall impact on the game. But does the holistic effect yield to an analytical representation, i.e. can it be expressed by means of independent evaluations of different parts of a position? Proceeding in a straight forward manner we would have to assume that all parts are homogeneous and have the same impact on the environment, i.e. they have the same weight (importance, value). Under these assumptions individual parts can simply be added together (two apples plus three apples equals five apples). The value arrived at represents one of the essential preconditions for further development. But suppose different parts are qualitatively different or, if they are the same, have different "weights." How then do we integrate them into a single whole? This can be accomplished if we take into account the "weights" of the parts (two apples weighing 80 grams each plus three apples weighing 100 grams each together weigh 460 grams; this characteristic is essential for further development). Determining (discovering) these weights is a rather difficult task. Physics has had the most success in this area. The situation with biological systems where these weights are represented by, say emotions, with economic systems where they assume the form of prices, and with a number of other fields is more complicated.[55]

So it is possible to express the holistic value of a given state through its parts analytically provided their integration proceeds by calculating their respective weights which reflect the marginal contribution of each part to the development of the whole. This is the way many mathematical procedures for solving optimization problems are structured. Condensed, global evaluations of different parts, namely the Lagrange multipliers generated by these procedures, are precisely these kinds of weights. They are global in a sense of allowing us to measure the marginal impact on the overall development of the system produced by independently varying one single variable at each local step. (Since chess pieces are discrete in nature and are rather small in number the idea of varying one single variable must be reconsidered). The global nature of Lagrange multipliers is due to the fact that the process by which they are formed takes all the constraints of the problem into account: the initial state, the final goal, and the rules of the "game".

The sum of the products of each component with its respective global value represents the level of system's development at a given stage; it measures the impact of the current state on the solution of the global problem.

If we look at chess from this perspective we can say that it is possible to give a global evaluation of a position in analytical terms at each local stage provided we know an exact algorithm of the game. This evaluation would reflect the impact of a given position on the outcome of the game. But what if such an algorithm is not available, what if there is no program linking parts with a whole in a rigid and precise fashion, how then do we determine the holistic effect of a given position?

May be we can try to apply the ideas of exact algorithms to determine approximate values of parameters used in evaluating the current state of the system. These values should be a good approximation of the weights generated by exact algorithms. Of course, it should be remembered that the problem of evaluating the overall position is the cornerstone of our analysis. Evaluation of position's individual parts or of a single piece remains an important issue as far as varying their values independently of each other, but it is essential to keep in mind the interrelationships between the parts and the whole.

What does all this mean? It means that we have to proceed from the general to the particular, i.e. from an overall position to individual pieces.

Two questions arise immediately: 1) how do we analyze the structure of a position, i.e. will it be sufficient to limit our information to such primary objects as pieces - material (together with their coordinates) or will we need additional primary objects? and 2) how do we measure the values of the pieces when we do not know their total worth relative to some final goal?

Affirmative response to the first question gives rise to two new questions which reveal the need to introduce new primary parameters to be used in the evaluation of a position. The two questions are 3) what is the role of this new set of primary parameters and 4) how to evaluate these new parameters so that they are commensurable not only among themselves but also with the weights of material objects which together provides us with a evaluation of the position as a whole?

Let us proceed to answer these questions.

The first question asks whether the information pertaining to such primary objects as material parameters is sufficient to evaluate a position. The answer to this question

---

[55] The difficulty of determining the exact relative weights of all the parameters lead to an attempt to simplify the problem in the following way: to try to discover one dominant part and use it to evaluate the impact of the entire state on the future. (In sciences, this idea is known as using one single characteristic to determine the overall state). But such dominant characteristics are rather rare.

bears directly on the second (as well as third and fourth) question which deals with evaluation of the material in inexact algorithms. I want to repeat that the structure of inexact algorithms rules out the possibility of linking the current state with the end of the game, i.e. of deducing the overall values of the pieces which would be sufficient to evaluate a position. We are thus forced to take a different route: first we assign unconditional values to the pieces, i.e. values corresponding to some ideal situation. We then try to introduce additional parameters which are supposed to reflect the current situation on the board, its specificity. This really represents an attempt to make up for the incompleteness of evaluations by including unconditional weights of the pieces. Now evaluation of these additional parameters which could be called compensatory parameters must account for their impact on the future flow of the game. Here parameters used to compensate for this deficiency pertain not only to individual pieces but also to groups of pieces or to the position as a whole. It seems to me that the nature of these parameters lies in the interrelations among the pieces in a given situation. It is precisely these relations which were not incorporated in the unconditional weights of the pieces. The more complete this set of parameters and the better their weight assignment the closer (other conditions being equal) is our evaluation of a position to that of an exact algorithm.

Therefore lack of information on the total value of the pieces (this information is needed to evaluate a position) can be compensated by determining the value of the entire position by means of unconditional values assigned to the pieces coupled with the parameters characterizing the relations among the pieces.

The next section explores the first component of the problem.

§ 2. Evaluation of material

Our preliminary discussion will focus on the analytical construction of unconditional values of the material. One reason for this is the availability of literature devoted to this topic. Another reason is the uniqueness, as I see it, of chess in its approach to determining the weight of the material.

Traditional treaties on chess - those devoted to the theory and practice of the game, are not concerned with the analytical methods of calculating the unconditional values of the pieces. For instance, E. Lasker's now classical chess manual does give us the unconditional weights of the chess pieces. The ratios presented in the above text are well known to all chess players. The author points out that these values are valid only in equal positions. The ratios between the values of the pieces can vary considerably depending on

the specific situation on the board. Just like many other players Lasker believes practical experience to be the source of the weights assigned to different pieces. The following quote from Lasker's book not only confirms my statement but also presents some possible relative numerical values of the pieces. These values will be important to us in the future.

> "...the masters, and among them certain regularities show very plainly. The experience derived from games played by such as deserve the title of masters, during centuries has proven these regularities."[56]

In his concluding remarks which really stand apart from the rest of the book Lasker gives us an interesting comparison between the values arrived at on the basis of practical experience and those produced by the mathematical analysis of the game.[57] He points out that Euler who was a great mathematician but a rather poor chess player was able to determine the values of the pieces mathematically and these values happened to correspond closely to the values representing the practical experience of great players.

The procedure used by Euler to determine the values of the pieces is not known. But Lasker points out that Euler's method "bears the seed of future discoveries"[58] Lasker does not give reference to Euler's work which could have been used to reconstruct his chain of reasoning using our present-day knowledge of the subject.[59]

Judging by Lasker's comments it is reasonable to assume that Euler used rather static positions to compute the values of the pieces, i.e. a case where material plays a decisive role. Two beginners playing in a random like fashion start from a given position and play n-games (n approaching infinity); material advantage for either side is determined by the winner of the match.

Lasker also notes that for Euler the mobility of a piece was probably the dominant characteristic in establishing its material value. Mobility, in turn, is determined by the free space or by the "obstacles" (thus a knight, although it is less mobile than the bishop, has an advantage over the bishop in being able to jump over the "barriers" set up by the opponent).

Well-known attempts at elaborating analytical methods to calculate values of the pieces were undertaken in non-chess literature.

Results derived from these non-chess sources coincide with practical experience to a surprising extent. This analysis is very valuable since it reveals the reasons why pieces have different weights. The weakness of these results lies in the fact that practical experience probably incorporates other important factors which may cancel each other.

---

[56] Lasker, E., Lasker's Manual of Chess, Philadelphia: Mc Kay Co, 1947, p. 35.

[57] Ibid., pp. 340-341.

[58] Lasker, E., Lasker's Chess Manual, Moscow: OGIZ, Fizkultura i Turizm, 1937, p.323.

[59] My own attempts to find this work were unsuccessful. I even asked specialists on Euler from the East German Academy of Sciences but they were unable to help me either.

The groundwork in the field of unconditional evaluation of the pieces (which happened to coincide with practical experience) was an article by H. Taylor.[60] In this work (I shall quote from it without further reference) Taylor proposed two methods of measuring the relative material values of the pieces. Common to both of these methods (this goes for Euler's method as well) was the dominant role of the material. Taylor wanted to reduce his experiment to the simplest possible case: the only pieces on the board were a king for one side and one piece for the other (besides king and pawns). Each piece wants to check the opponent's king. This is the set up of the problem:

"A king and a piece of different colors are placed at random on two squares of a chessboard of n^2 squares: it is required to find the chance that the king is in check."

Unlike Euler's experiment, this method of calculating the relative values of the pieces for a given size board is based on a finite number of trials since the number of squares that can be occupied by a king and one piece is limited by the size of the board.

Taylor discovered that his method of calculating relative weights of the pieces produces results different from those used by the players. So he modified his experiment. Instead of calculating the probabilities of king checks he introduced an additional requirement of a "safe check." "Safe check" differs from a "simple check" in that the piece making the check cannot be captured by the opponent's king. The notion of a "safe check" incorporates two aspects: the mobility of the piece plus its ability to "overcome obstacles." In order to determine the relative weights of the pieces which E. Lasker spoke about Taylor elaborated a formula to calculate the probability of a king being in check, both "safe" and "simple." Table 3 below is borrowed from Taylor's article. It shows the results of his experiment for any size board having an even number of squares and for a regular board having 64 squares. Here, n stands for the number of squares on the board.

Ordering (reducing to a common denominator) the probabilities of different pieces putting a king in check gives us the following sequence: for the case of a "simple check" the values are 3, 5, 8, and 13 for the knight, the bishop, the rook, and the queen respectively; for the case of a "safe check" they are 3, 3.25, 6, 9.25 (12, 13, 24, 37). The correspondence of the last set of figures to the ratios commonly used by the players (they are 3, 3, 5, 9) is amazing.

Table 3. The probability of a king being in check

	Chess board with n2 squares		Regular board	
	Normal	Safe	Normal	Safe
Knight	8(n-2)/n2(n+1)	8(n-2)/n2(n+1)	1/12	1/12
Bishop	2/3x2n-1/n (n+1)	2/3x(n-2)(2n-3)n (n+1)	5/36	12/144
Rook	2/n+1	2(n-2)/n(n+1)	2/9	1/6
Queen	2/3x5n-1/n(n+1)	2/3x(n-2)(5n-3)n (n+1)	13/36	37/144

Working within an even more general setting M. Kavesh suggested evaluating a pawn by the number of safe checks it can give from the eighth rank where it can be promoted to a queen. The results coincided with the established value of the pawn relative to other pieces.

Attempts independent of those by Euler and Taylor to determine the relative values of the pieces analytically were made in more recent times. The basic idea was to have a clear board with one single piece. Then an average number of moves that a piece can make depending on its position on the board was calculated. For instance, a rook can make 14 moves from any square on the board. Obviously, an average number of moves it can make is also 14 (14x64/64). A queen can make 21 moves from 28 squares of the board, 23 moves from 20 squares, 25 moves from 12 squares, and 27 moves from 4 squares. Altogether a queen can make 1456 moves (21x28+23x20+25x12+27x4). Its average mobility is 22.75 moves (1456/64). Similar calculations can be performed not only for the knight and the bishop but also for the pawn and the king. Thus, a knight can make an average of 5.25 moves ((2x4+3x8+4x20+6x16+8x16)/64), a bishop can make 8.75 moves ((7x14+9x10+11x6+13x2)/32). As an autonomous piece, a king can make 6.5625 moves on the average ((3x4+5x24+8x36)/64); a pawn can make an average of 2.5 moves ((2x10+3x32+4x6)/48). Here, it is assumed that a pawn cannot occupy the first rank or move backwards from the eighth rank and the possibility of its promotion upon it reaching the eighth rank is ignored.

Below we have a formula for calculating an average number of possible moves. They are denoted by P_n for all the pieces[61] and the size of the board which is arbitrary is denoted by n.

---

[61] Derivation of the formula for the bishop is rather complicated so Gik recommends a book by Okunev, L. J. entitled Combinatorial problems on the chess board, Moscow: ONTI, 1935.

$P_n(K)=4(n-1)(2n-1)/n^2$
$P_n(Q)=2(n-1)(5n-1)/3n$
$P_n(R)=2(n-1)$
$P_n(B)=2(n-1)(2n-1)/3n$
$P_n(Kn)=8(n-1)(n-2)/n^2$
$P_n(P)=(n-1)(3n-4)/(n-2)$

Taking pawn as a basic unit of measurement the relative values of the pieces are 2.1 for the knight, 3.5 for the bishop, 5.6 for the rook, and 9.1 for the queen. These ratios in the values of the pieces are very close to those worked out by Taylor who used the "simple check" criterion (values arrived at by Taylor must be divided by 1.428 to even them out with the ones generated by the above method). This fact is not a coincidence. Indeed, the formula for king checks differs from the formula for the average number of moves that can be made by a given piece in only one respect: the former method takes the average number of possible moves available to a given piece relative to the total number of squares on the board and the latter takes the sum total of all possible moves and divides it by the total number of squares on the board.

An example with a rook should help illustrate the point. Using Taylor's method (the "simple check" criterion), the number of checks that a rook can make is 14 no matter what the position of the king is, i.e. $2(n-1)$. On the average, a rook will be able to do so from (n^2-1) squares since one of the squares is always occupied by itself. Therefore, the probability of a rook check is $2(n-1)/(n^2-1)=2/(n+1)$. This formula differs from the one which calculates the number of possible moves $2(n^2-1)/(n+1)=2(n-1)$ by a multiple of (n^2-1). In other words, in calculating the overall probability of a check that a rook can make from a given square we discover that it is precisely $2(n-1)$. The same considerations apply to cases where the number of possible checks varies depending on the position of the piece. A reader can perform these exercises by putting (n^2-1) in the numerator of each of the formulas. A discrepancy between the value of (n^2-1) and 1.428 - a coefficient used to compare the values of the pieces calculated according to one of the aforementioned methods, arises as a result of the chosen standard (a particular unit to measure the relative values of the pieces).

Although we arrive at the same relative values using either the Taylor's formula for "simple checks" or the mobility of the piece formula, the assumptions inherent in these two approaches have very different implications. Taylor's formulation of the problem is more meaningful: it relates the value of the piece with the final objective of the game. This simplifies our task of finding the appropriate set of constraints to be incorporated into our experiment aimed at establishing the relative weights of the pieces and make our analytic results resemble those arrived at on the basis of practical experience. The other more abstract approach which is based on piece mobility does not guarantee affinity of the results. Still, it is advantageous in some respects; for instance, it allows both the king and the pawns to be evaluated as active pieces.

To sum up we can state the following: the relative values of the pieces are functional in that they reflect the potential contribution of a given piece to the final goal of the game. At the same time, these values are idealized, i.e. they are deduced under "perfect" conditions practically unattainable in the real game. These values are rather unusual being averaged out and completely unconditional. The choice of the averaged out unconditional values rather than the highest ones generated by the most favorable circumstances "softens" the unconditional nature of these values. It makes the game more concrete. For instance, it takes into account the possibility that in some distant future a king and one piece may come to be positioned on any squares of the board.

This approach to evaluating the pieces leads to conclusion that the value of a piece is always the same. To make its value more concrete, a piece has to be examined within the context of a specific position. This is where positional parameters come in. For instance, some players claim that the values of the pieces in the opening are different from the usual rating. This means that the values of the pieces change as a result of positional considerations. E. Lasker writes that in the opening the relative values of the pieces differ depending on their initial positions.[62] They also differ from the commonly accepted weights of the pieces. The second column of Table 4 gives the values of the pieces according to Lasker; other columns compare his values with the commonly accepted ones.

Unfortunately, Lasker does not discuss any analytical methods for generating the values of the pieces in the opening. Still, it is easy to see that one of the major factors responsible for making the opening values of the pieces different from their overall ones is the immobility of the pieces in the opening, the fact that some pieces limit the development of others.

It would be very interesting to do an experiment to represent the modified (Lasker's) values of the pieces in the opening analytically by introducing a number of positional parameters. I leave this exercise to the reader.

---

[62] Lasker, E., Lasker's Manual of Chess, Philadelphia: Mc Kay Co. 1947, pp.106-107.

Table 4. A comparison of the values of the pieces in the opening (suggested by Lasker) with the established values.

The piece	Its value in the opening	Averaged out value: opening, by type of piece	Averaged out value: opening, rel. to a pawn	Averaged out value: general	The difference between relative averaged out value in the opening and overall: absol.	The difference between relative averaged out value in the opening and overall: %
Pawn: d or e	2	1.3	1	1	0	0
" c or f	1.5					
" b or g	5/4					
" a or h	1/2					
Knight	4.5	4.5	3.5	3	+0.5	+16.7
King bishop	5	4.75	3.65	3	+0.65	+21.6
Queen bishop	4.5					
King rook	7	6.5	5	4.5	+0.5	+11.1
Queen rook	6					
Queen	11	11	8.5	9	-0.5	-0.6

Concluding our discussion on evaluating material parameters, I want to repeat that our ability to deduce these values is largely due to the peculiar set up of the game discussed in the first part of the book. The only point to add to that discussion is that every piece in chess except a king can pose a direct threat to enemy king even if the piece itself is "not in danger," i.e. every object in chess has the capacity to be directly involved in achieving the final goal. This applies to pawns as well for they can be promoted to stronger pieces enabling them to fulfill the above mentioned function.[63]

§ 3. The structure of positional parameters

I hope the above analysis of the methods of forming unconditional material values has made it clear that other parameters have to be introduced in order to improve upon our unconditional values in each specific situation. Doing so will enable us to give an overall evaluation of the current position (or some possible future position). These additional parameters must reflect interrelations among the pieces. As I have already mentioned above the purpose of introducing these parameters is to do explicitly what an exact algorithm of the game does implicitly.

I want to start our discussion with primary positional parameters, i.e. parameters that cannot be decomposed any further with the tools presently available. Primary parameters can be grouped together to form aggregate positional parameters such as the strength of the center, the development of the right or the left flanks, and so on. I shall not touch upon this kind of parameters in the present work. I only want to note that a lack of research on the methods of aggregating-disaggregating chess positions represents a gaping hole in the field of computer chess. There is no doubt that good chess players utilize these kinds of procedures to a great extent.

---

[63] All this leads to a very interesting question which still remains open: "To what extent can the methods used in chess to form unconditional weights of the pieces be used as a general principle in other systems having some universal intermediate objects, i.e. objects not directly related to the final goal even if such a goal does exist?"

Beginning with Steinitz chess players have paid a lot of attention to primary positional parameters. Examples of these parameters are scattered throughout E. Lasker's chess manual. For instance, on page 202 he examines such positional parameters as double pawns, backward pawns, and isolated pawns. Lasker also mentioned a number of positional parameters that have not received wide acceptance till this day. He stressed the importance of positional parameters usually associated with pawn structure in analyzing piece coordination. Elaborating upon Steinitz's ideas, Lasker points out

> "If, one the one hand, the distance of two pieces, for instance of two knights, is very great, a co-operation or a conflict between them is possible only in the future, unless they co-operate or contend with each other indirectly..."[64]

It seems that the first systematic list of positional parameters was put together by C. Shannon.[65] As far as I know, it is still the only systematic list of positional parameters to be found in the chess literature; therefore, it is given here in its entirety.

"(2) Pawn formation:

(a) Backward, isolated and doubled pawns.

(b) Relative control of center (pawns at K4, Q4, B4).

(c) Weakness of pawns near king (e.g. advanced KNP)

(d) Pawns on opposite color squares from bishop.

(e) Passed pawns.

(3) Positions of pieces:

(a) Advanced knight (at K5, Q5, B5, K6, Q6, B6), especially if protected by pawn and free from pawn attack.

(b) Rook on open file, or semi-open file.

(c) Rook on seventh rank.

(d) Doubled rooks.

(4) Commitments, attacks and options:

---

[64] Lasker, E., Lasker's Manual of Chess, Philadelphia: Mc Kay Co, 1947, p.239.

[65] Shannon, G., "Programming a Computer for Playing Chess", The Philosophical Magazine, vol.XLI, 1950.

(a) Pieces which are required from guarding functions and, therefore, commited and with limited mobility.

(b) Attacks on pieces which give one player an option of exchanging.

(c) Attacks on squares adjacent to king.

(d) Pins. We mean here immobilizing pins where the pinned piece is of value not greater than the pinning piece; for example, a knight pinned by a bishop.

(5) Mobility." (p. 274).

Authors of chess program "Caissa" came up with a more elaborate but still not systematized set of positional parameters. The first version of the program included 18 parameters.[66] Subsequently the list was increased to 31.[67] Classification of the parameters included in this list is rather simple. First of all, 12 parameters associated with squares and pawns were distinguished from those associated with squares and pieces. The authors had also come up with 10 parameters reflecting the dynamics of the game. Among these parameters were "attacking the opponent's strongest piece," "a light piece attacking a weak square in the opponent's position," etc.

This is what the authors wrote concerning the subject:

"...at first, a rook must attack a square lying on an open file, then it must occupy it, and finally "invade the gluttony row". The opponent, in turn, must try to prevent this from happening by creating a position devoid of such weaknesses."[68]

And finally they distinguished another very specific parameter "Bishop attacking any piece belonging to the opponent (except a pawn)" which is really a particular case of another parameter. A parameter called "Nail" which describes a particular piece and pawn formation was also introduced.

I want to propose another classification of positional parameters presented in the aforementioned works on the "Caissa" chess program. It is given in Table 5. I believe this classification is more systematic. Its matrix representation reveals the connections between the relations and the objects comprising them. Pawns, because they are rather special (for instance, they have less mobility than other pieces and form relatively static structures - the skeleton of the game), are allotted a special place in this classification.

Table 5. Classification of positional parameters

Types of the relations	Object of the relations					
	Squares	Pawns	Squares and pawns	Pieces	Piece and pawns	Squares,Pieces and pawns
Configuration	12.Open file****	2. Phalanx 6. Isolated Pawn	3.Center pawn* 5. Passed pawn 9.Isolated pawn on a semi-open file*****	32. 'Nail^^^^ Pin^^^^^	19. Light piece occupying opponent's weak square 20. Knight posted in the center 22. Queen (rook) on an open or a semi-open file 23. White rook on the seventh or eight rank++++	(2)(c) weakness of the pawns near kingΔ
Control	10. Weak squar**		4. Pawn attacking a center square	14. Scope + 15.Opponent's highest ranking piece under attack++ 16. Opponent's unprotected piece under attack+++17. Fork++++ 31. Bishop attacking any opponent's piece (except a pawn)	18. Light piece hitting opponent's weak square 21. Rook (queen) on an open or a semi-open file 24. A square next to the opponent's king under attack 25. Queen-king scope++++++ (4)(b) attacking pieces with possibility of exchange	
Quantity and variety	number of free squares					(3)(g) pair of rooks^^^ variety of pieces bishop&rook vs. knight&rook; queen&knight vs. queen&bishop; two bishops vs. two knights^^^^
Loosing opportunity to perform a special operation					27. King castled 28. King-side castle given up° 29. Queen-side castle given up°° 30. Castle given up°°°°	

---

[66] Adelson-Velskii, G., et all, "Programming the Game of Chess", Soviet Mathematical Proceedings, v. XXY, Issue 2/152, 1970.

[67] Adelson-Velskii, G., et all, Machine Plays the Game, Moscow: Nauka, 1983.

[68] Ibi

Comments to Table 5.
Arabic numerals refer to the list of parameters taken from the book by Adelson-Velskii, G., et
all,1983, Machine Plays the Game. Moscow: Nauka. If the Arabic numeral is accompanied by a
letter index and both are in parenthesis, then they refer to the parameters taken from the 1950
work by C. Shannon. No number in front of the parameter means that it was borrowed from other
sources referred to separately.
Upper indices indicate that the respective parameter is accompanied by a footnote; I introduced
all these symbols only for the sake of convenience. The need for explanations arose because
players and programmers sometimes use different terms for the same parameter and some
positional parameters are expressed in chess jargon which may be unfamiliar to the reader.
Explanation of the parameters:
*For white - squares e4, e5,e6, d4, d5, d6; for black - squares e5, e4, e3, d5, d4, d3.
***(Hole) - a square under attack by an opponent's pawn and is not and cannot be defended by
ones own pawns unless it moves to a different file as a result of a capture.)" -
****(Weak pawn) - a pawn standing on a square that lies behind a "hole".
****Open file
*****File open to a given player - only opponent's pawn(s) on the file
+ "For each piece, including a pawn, the squares that come under its attack including those
occupied by its own pieces or pawns (squares it defends)."
++"A piece occupying a square that is under attack by an opponent's piece(s) but not defended
by one's own piece(s)."
+++"A piece occupying a square that is under attack by an opponent's piece(s) but not defended
by one's own piece(s)."
++++"(Fork) - opponent's piece attacking two undefended higher ranking pieces of a given
player."
+++++ Black rook on the first or siccing rank.
++++++ "A square that can be occupied by a queen if it was placed on a square currently
occupied by the king. Having occupied it, opponent's queen (from vertical and horizontal squares
also a rook, bishop from the diagonal squares) gives your king a check."
° A rook occupying the h1 square in the starting position has moved
°° A rook occupying the a1 square in the starting position has moved
°°° Both rooks or a king have moved
°°°° "Any piece or a pawn in front of the opponent's pawn occupying its original position
between the c and the f files. This structure splits the flanks of the opponent."
°°°°° A given player must protect the piece in question with another piece
^ For instance, an advanced king night pawn
^^ Here, they mean a "pin" where the value of the pinned piece does not exceed the pinner; for
instance, a bishop pinning a knight.
^^^ Variety of pieces, especially important in combinational play since it creates more
opportunities.
^^^^ According to Capablanca, a bishop and a rook are stronger than a knight and a rook, but a
knight and a queen may turn out to be stronger than a bishop and a queen; two bishops are
almost always stronger than two knights.[69]

I believe by now the reader has sufficient empirical material to understand the diversity of positional parameters.

---

[69] Capablanca, H-R., Chess Manual. Moscow: Fizkultura i Sport,1975, pp. 30-31.

I want to make one additional remark. It is very unclear whether the above set of positional parameters is exhaustive. First of all, the language of the chess players is rich but it lacks precision. Therefore, just the formalization of parameters in chess players' arsenal presents certain difficulties: even qualified chess programmers are sometimes unable to do this.[70]

Assuming these difficulties can be overcome, there are other more serious problems. We can look at objects from an infinite number of perspectives so the number of positional parameters also seems to be infinite. Choosing a subset of positional parameters relevant to a given problem still remains an unresolved issue.

It is close to impossible to construct a system from which we can deduce the entire set of positional parameters which would be sufficient to create an "invincible" algorithm of the game.

Unfortunately, the set of positional parameters currently used by computer programs is still based on the experience of good players who express it in an intuitive manner. I think that one of the major obstacles to designing more effective algorithms is a lack of formalized procedures for developing new positional parameters. These procedures lie at the foundation of second-level programs, i.e. programs that modify other programs directly responsible for generating the moves.

Finally, let us consider the problem of evaluating positional parameters since their share in the overall value of a position may vary. I know of no analytical works on this subject that would even resemble analogous works on the evaluation of material parameters. In fact, I am familiar with two works which contain intuitive evaluations of positional parameters worked out by expert players.[71]

Whereas the first work presents only intuitive evaluations of positional parameters, the second one elaborates a number of ideas on how to form them. In Slate's work the general idea of assigning values to positional parameters (as well as material ones) is founded upon the following assumption:

> "Basically, the function adds several easily computable factors together. The dominating term is material. The remaining terms are rules of thumb designed to give the program a vague positional sense. Rather then expressing some theoretically

---

[70] An interested reader can consult the book by Adelson-Velskii, G., 1983, pp. 59-60, which discusses this aspect in detail.

[71] Adelson-Velskii, G., et.all, Machine Plays the Game. Moscow: Nauka,1983. (The last book by the same authors which provided material on positional parameters does not contain this kind of evaluation; and Slate, D., "The Northwestern University Chess Program", Chess Skill in Man and Machine. New York: Springer Verlag, 1977.

rigorous chess principles, these factors merely improve somewhat over 'aimless' moves. The design motive for the evaluator appears to be a self-fulfilling prophecy: we knew that it was going to be primitive and would thus depend heavily on tree searching. To do such deep tree searching would require that the evaluator do its work quickly, and so it wouldn't have the time to do anything clever, and thus would have to be primitive."[72]

The work quoted above encouraged others to try to determine at least partially (with a large number of parameters assigned intuitive values) the values of a few positional parameters analytically.

For instance, "Knights are evaluated and king and center tropism and development. Center tropism is a peculiar measure of the closeness of a knight to the center of the board and is computed as 6 minus twice the sum of the rank and file distances from the center of the board to the square the knight is on. The value is then multiplied by 1.6...

King tropism, for knights, is computed differently from that of rooks. We compute 5 minus the sum of the rank and file distances between the knight and enemy king and multiplied the result by 1.2... Like bishops, development of knights means not being on the back rank, the penalty being 9.4 ." (p.97).

Apparently, the weights of a number of positional parameters were not formed analytically: they were based primarily on players' intuition and experience.

Reasons underlying this method of evaluating positional parameters are deep. Subjectivity of judgement is rooted in the indeterminate nature of the game, in the need to adjust the play to the abilities and knowledge of a given player. Analysis of material parameters has revealed that they arose as as result of the rules of the game itself and their weights were calculated in an objective fashion. In general, subjective evaluation of a state is due to the subjective set of chosen positional parameters and their subjective evaluation. These considerations regarding the subjectivity of positional parameters do not rule out the possibility of establishing some averaged out magnitudes based on the analysis of a large number of games by players or machines. Still, these averaged out values would be mere guidelines, i.e they will reveal at first glance parameters which must be taken into account and call attention to those parameters whose subjective evaluations differ substantially from their averaged out norm; reasons for the gaps would have to be explained in order to justify the use of these averaged out values.[73]

M. Botvinnik proposed a totally different approach to positional parameters.

---

[72] Slate,D., and Atkin,L., " The Northwestern University Chess Program", Chess Skill In Man and Machine. New York: Springer-Verlag, 1977, pp. 93-94.

[73] This question is investigated in detail in the forthcoming book "Indeterministic Economics" with regards to evaluating the state of the firm.

Acknowledging the need to evaluate both material and positional parameters, he notes that:

"Such a method (method by which various averaged out weights of positional parameters are introduced A.K.) of deriving a positional estimate in chess is wrong. A positional factor that yields a positive result in one situation may yield a negative result in another. For instance, are doubled pawns good or bad? The answer depends on the situation. Sometimes doubled pawns are a suitable target for an attack, since one of them cannot support the other. But sometimes doubled pawns contribute to the control of a square through which important communication lines (trajectories of pieces) pass, and then the doubling is extremely useful. The same thing may be said of other factors entering into the positional component of Shannon's prototypical scoring function.

In the chess program under consideration, a quite different decision was made about the positional component of the scoring function, basing it on the control of those squares making up the trajectories that enter the MM. The side controlling the larger number of squares has a positional preponderance."[74] (p.29).

Botvinnik elaborates on this point:

"Control of fields does not mean control of the whole board, but control of only those fields that may be used in the impending play. Therefore one must strive for control of the field consisting of those trajectories in which the pieces can move, but have not moved yet.

At the node in the search tree where we find ourselves at a given moment, we must unravel all those sheaves of trajectories which have not yet been developed and determine which player has control of the majority of the fields consisting of the trajectories not yet used in the play. This allows us to forecast the result of the play - - the result of a search which, in particular, had to be renounced at the terminal nodes of the variations for lack of resources."[75]

An interested reader can refer to Botvinnik's book quoted above for a more detailed discussion of the methods of assigning weights to positional parameters. I do not intend to compare the two approaches to the evaluation of a position proposed by M. Botvinnik and C. Shannon (or his followers). Again, I refer the reader to G. Adelson-Velskii's book which presents a comparative analysis of Botvinnik's approach.

Finally, evaluation of a state calls for an answer to the following question: "How do we compare the value of material parameters with the value of positional parameters?" The answer given in the literature is very intuitive. Essentially, it claims the the total value of positional parameters should be somewhere between that of a half-pawn to two pawns. Adelson-Velskii notes that the value of the smallest material unit - a half-pawn "must be high enough to exceed the total values of all positional parameters."[76] At the same time, they say "It would be very beneficial to think of positional parameters comparable to the material in value."[77] D. Slate, D. writes "The total of the nonmaterial terms does not usually exceed the value of about a pawn and a half."[78] And finally a quote from M. Botvinnik: "We may assume that a so-called positional sacrifice should not exceed two units of material (this will be made more precise in an experiment)."[79]

---

[74] Botvinnik, M., Solution of Inexact Problems. Moscow: Soviet Radio,1979, p. 29.

[75] Ibid., p. 38.

H. Berliner notes:

> "I differ from some leading practitioners in that I have always been an advocate of positional contributions that can some to more than the value of a Pawn. We have had situations where Hitech considered one side to have almost 3 pawns worth of positional compensation."[80]

This emphasis on increasing the relative weight of positional parameters stems from the experience of professional players. In some cases, they are willing to sacrifice material and not just a pawn but a whole piece in order to improve their position. But here we are jumping ahead of ourselves. I will have a chance to discuss positional sacrifice when I talk about the relationship between combinational and positional styles of play.

---

[76] Adelson-Velskii, G., et all, Machine Plays the Game, 1983, p. 69.

[77] Ibid.

[78] Slate, D., and Atkin, L., " The Northwestern University Chess Program", Chess Skill in Man and Machine, New York: Springer-Verlag, 1977, p. 95.

[79] Botvinnik, M.M., Computers in Chess: Solving Inexact Search Problems, Springer-Verlag: NewYork, 1984, p. 38.

[80] Berliner, H., "Some Innovations by Hitech", ICCA Journal, Vol. 10, No. 3, 1987, pp. 111-117.

Chapter 4. Class of local problems and stages of the game.

§ 1. Classification of search problems

Having better understood the evaluation of a position we can discuss different types of local search problems. Problem type depends primarily on the structure and the value of the final product associated with each local problem. What determines the constraints of the problem is the current position on the board.[81]

The structure of the final product can be represented in a variety of ways. Our classification is based on four characteristics.

1. How fixed are the values of independent input variables, i.e. some or all of the variables are given as constants or as unknowns; all this is based on the hypothesis of the existence of the true values of the variables incorporated into the objective function.

2. Remoteness (number of moves separating) of the final position from the current one: the final position (relative to a given problem) can be distant or close (even one half-move away) to the current one.

3. The extent to which the structure of the objective function is certain: the qualitative characteristics of its arguments may or may not be well defined.

Table 6. The structure of the objective function of the local problem.

Characteristics of the future positon		Remoteness of the future position from the current one			
		Far		Near	
		Impact on the development		Impact on the development	
The values of the variables determining how fixed is the value of the position	The degree of uncertainty of the structure of the position	strong	weak	strong	weak
Constants	Clear				
	Fuzzy				
Unknown	Clear				
	Fuzzy				

---

[81] Of course, this is only true if we assume that the process is markovian. Actually, it may be worth while in formulating local problems to remember previous moves made in the game. Here, I am going to ignore these considerations.

4. The variability of the value of a given parameter as a function of the impact of a current position on the future flow of the game.

Table6 presents all possible combinations of these characteristics in amatrix form.

Table 7 gives names and abbreviations for the 16 types of local problems derived in Table 6.

Table7. Types of local problems and areas of their application

No. Area of application

1. Positional problem with a highly effective procedure for eliminating inferior moves
2. Positional problem; the procedure for eliminating inferior moves is not very developed
3. Positional problem given in aggregate variables with a highly effective procedure for eliminating inferior moves
4. Positional problem given in aggregate parameters; the procedure for eliminating inferior moves is not very developed
5. Sublocal problem of the general form with an unexpectedly strong outcome
6. Non-key sublocal problem of the general form (standard algorithm block)
7. Sublocal problem given in aggregate parameters with an unexpectedly strong result
8. Sublocal problem given in aggregate parameters with a weak result
9. Positional goal directed type of problem with sublocal problems
10. Goal directed type of problem with sublocal problems, result weak
11. Goal directed type of problem given in aggregate parameters with sublocal problems; produces drastic changes on the board
12. Goal directed type of problem given in aggregate parameters with sublocal problems; weak results
13. Combinational problem; the amount of material captured is decisive
14. Non-key sublocal problem
15. x
16. Combinational problem; material is won and positional parameters vary.

I want to introduce based on this classification some general definitions pertaining to local problems. If parameters associated with the final position are given (they are known values) this position shall be called a goal.

A problem based on this kind of a goal shall be called a goal-directed problem.

If the number of moves separating the final position from the current one is more than one the problem shall be called protractive; a one-step problem shall be called a condensed problem.

Suppose all the initial as well as final conditions of a goal-directed protractive problem are given. Then the problem reduces to finding an optimal trajectory of moving from one state to the other. This trajectory could, for example, fulfill the requirement of reaching the final state in a minimum amount of time. A function designed to implement this kind of a constraint shall be termed a criterion of optimality. In an extreme case when the choice of the trajectory does not make any difference a local optimization problem reduces to finding just one such trajectory which of course must not violate the rules governing the transformation of one state into another.

Unlike goal-directed local problems local problems that include parameters of the final state whose values are unknown shall be called non-goal-directed problems. Here, a function which evaluates a final position shall be called an weight function. Non-goal-directed problems can be either protractive or condensed.

In setting up a local problem we may have some idea of the value of some part of the final position in question. Our hypothesis does not have to be incorporated into the evaluation of the final position explicitly; it can be represented as a constraint assigned a constant value. These types of constraints represent objective constraints. They should not be confused with global constraints which also stay fixed but which pertain to the local problem in its global form; these constraints are uncontrollable, they are completely determined by previous development.

All the terms associated with the definition of a final position (goal, objective constraints, criterion of optimality, weight function) combine to form the concept of an objective function. It allows us to focus on the general structure of the solution of the local problem in case we are not interested in all its details.

§ 2. Non-goal directed protractive problems.

Let us begin our analysis with non-goal directed type of local problems.

Constraints incorporated into local non-goal directed condensed problems are determined by the current position of the pieces (both one's own and the opponent's). The value of a move is determined by the weight function as the difference between the evaluation of the new position (one half-move) by a given player and his evaluation of opponent's half-move. It is assumed that the opponent responds to your own best half-move with his best half-move. This method of evaluation is based on Zermelo's principle. It states that the value of a move is determined by the maximum value of a position reachable by that move.

All this is rather trivial. It is analogous to a firm trying to maximize its profit or a man trying to maximize the difference between his positive and negative emotions. (It is also akin to the process of setting up local problems in the above mentioned decompositional algorithms of optimization).

In constructing a weight function for a given problem it is assumed that the values of all material and positional parameters included in it are unknown, i.e. all its arguments are variables or unknown quantities. In other words, we cannot predict the exact values of all the parameters determining the value of the position at the end of the search; i.e. we do not know which pieces shall stay on the board, what squares they will occupy, or which positional parameters associated with that particular configuration will come into play.

At the same time, it is assumed that all the coefficients of the arguments, i.e. weights of material and positional parameters are completely determined. The other two components of the objective function cannot be determined outside a given situation. For instance, the "distance" separating the final position from the current one largely depends on the search algorithm being used, the ability to eliminate inferior moves, and the power of the computer. The impact of unknown structure on the future of the game is usually rather weak, i.e it is not meant to accommodate major changes. At the same time, it can only be determined post factum or after the problem has been solved.

If we knew the overall values of parameters included in the local objective function, i.e. values conveying global information about the entire game, then the solution of such a one-step local problem would constitute one segment of the optimal trajectory representing the solution of the global problem. But as I have mentioned many times before such global evaluations cannot be deduced within an algorithm of an inexact game: we have to settle for partial evaluations. Therefore the value of the local objective function is not exact. This has lead many people involved in the development of chess algorithms to expand the search horizon, i.e. increase the depth of the search. It is clear that the lesser the depth of the search the greater the probability, all other conditions being equal, of undiscovered flaws in the positions beyond the horizon: evaluation of a position at a given step serves merely as a guideline for future play.

Now, since the values of the variables comprising the objective function are not known, the local problem at the start of the procedure would, in theory, involve just one move - two half-moves. The actual number of moves looked at, i.e. the horizon of the protractive problem depends on several aspects including the power of the algorithm and the machine being used. It was precisely this uncertainty over the length of the sequence of local one-move problems that stimulated me to call this sequence of local problems a protractive local non-goal directed problem.

As I have noted in the section on general description of a chess algorithm, the number of possible moves in the game is enormous so the machine is really limited in its ability to exhaust all the variations. Therefore, the solution involves a well-known trade off between the depth of search or the horizon and the width of search or the number of moves examined at each step.

It was precisely this kind of approach to the formulation of a local problem and a corresponding objective function (viewed within the global framework of the entire algorithm) that was suggested by Shannon, C., 1950. This approach, somewhat modified, is still being used today by most algorithms of the game.

At the same time, there are strong arguments against Shannon's approach. M. Botvinnik thinks that

> "The first method, however, is hopeless as means of finding a good solution to an inexact task, as we shall show by applying it to the game of chess. On the average, each side in a chess game has 20 available moves. Suppose we start in a position where White is to move; all 20 moves of White's pieces must be included in the search tree. Then it is Black's turn; for each of White's 20 moves, Black has 20 answers, so that the tree contains 400 Black moves. If we extend the variation to include another move by White, the tree will complete moves -- six half-moves ("plies") -- it will contain some 67,000,000 nodes! The size of the tree is an exponential function of the depth.
>
> If we develop a suitable algorithm for a chess program, the resources of the computer (memory and speed) are less significant than might be supposed at first glance. Whenever we extend the depth of the variation by a move, the resources of the machine must increase by several factors of 10; this is impossible in practice. Thus the main element is not the power of the computer, but the skill of the programmer, i.e., the extent to which he has contrived to depart from the fundamental principle (a full-width search) in order to economize on machine resources."[82]

I agree with Botvinnik's criticism of protractive problems but I do not think they should be ignored altogether. Not only is the protractive approach capable of competing with other approaches in formulating local problems, but it can also be used in conjunction with other techniques. I shall elaborate on this idea further on in the chapter.

§ 3. Goal-directed protractive problems

In the present section, rather than concern myself with the ways of improving the set up and the solution of non-goal directed protractive problems, I shall try to apply other principles to resolve these problems, principles rooted in the goal-oriented approach especially its protractive variety.

In any goal-directed problem it is assumed that the goal can be reached without breaking the rules of the game. A reader may be surprised by this seemingly naive and indispensable requirement. But there is a branch of chess theory called retrograde chess.[83] There are positions that are legal as far as the positions of the pieces are concerned but which can never be reached from the starting position if we follow the rules of the game. The purpose of retrograde analysis is to prove that there exists no backward sequence of legal moves from a given position (legal in itself) that would eventually lead back to the starting position.

Therefore, when we form a long-term goal we have to be careful to avoid "retrograde goal".

Assume there is no risk of such a goal being set (it does not occur frequently anyway so we can ignore it for the time being). Still we have to set a goal that is realistic, not just wishful thinking on our part (that is random in the extreme case). So far, there is no formal procedure to accomplish this task. I shall attempt to show below that the solution to this problem calls for a rather sophisticated analysis of the structure of the state (current position) - uncovering positional features peculiar to it. This would ensure that the goal set is more or less reasonable.

We shall begin our analysis with attractor type of problems. Its impact on the outcome of the game is strong and its specific goal is given over a limited set of material parameters (problem number 13). A typical goal for this kind of problem is to try to capture opponent's material which would virtually defeat his position.

In other words, positional considerations can be ignored when the goal is to capture a lot of enemy material. The criterion of the local problem reduces to finding an optimal trajectory linking the current position with the final state to be attained.

It is customary to call this kind of problem combinational.

---

[83] Smullyan, R., The Chess Mysteries of Sherlock Holmes. Hutchinson, 1980. This branch of chess, sometimes called "chess logic" is also becoming part of computer chess. See Alden, B. and Bramer, M., "Development of a Problem for Solving Retrograde Analysis Chess Problems", Advances in Computer Chess, 3, ed. by M. R. B. Clarke, Oxford: Pergamon Press, 1982, pp. 121-137.

The section entitled "Forced play and how to prepare for it" of Adelson-Velskii's book[84] quoted many times presents some interesting examples of a goal (only in simple cases like mate in one) which consists of capturing enemy material without searching through all the variations branching from the current position.

It is easy to see that combinational problems have some very profound advantages: its solution gives one side an upper hand by generating a swift attack aimed at winning enemy material.

Nevertheless these advantages are rather conditional, they are realizable only under some very special circumstances. Although the goal associated with a combinational type problem is relatively clear, it suffers from a number of drawbacks which stem from the need to determine the values of different pieces. Of course, if one side is up a heavy piece then the evaluation of a position is generally not a problem. Here, "generally" excludes, among other things, an intentional sacrifice on the part of the opponent. (The subject of a sacrifice in chess will be addressed in the next chapter.)

But what happens if the consequences of a combination are unclear or if we cannot come up with such a combination at all?

> "Maybe a combination does not exist at all, may be he [player] is just chasing a ghost? A search for a combination may turn out to be futile. To base one's play on the existence of a combination is a very dangerous enterprise."[85]

How then is the game to be played? It must be played positionally. Positional play does not preclude combinations. On the contrary, it creates predisposition towards a combination. To quote Lasker once again

> "Hence, he [Steinitz] concluded that some characteristic, a quality of a given position must exist that to a discerning eye would indicate the success or the failure of the search before it was actually undertaken. And this characteristic, if explicable by reason, of what could it possibly consist if not an advantage or a disadvantage...And an advantage, if reasonable, what could that be except the thing that was generally termed so: greater material force, greater mobility, greater effect against the King..."[86]

So positional play makes sense if it eventually leads to a combination. To rely exclusively on combinations is flawed for it rings of primitive pragmatism. Purely positional style has the disadvantage of being over speculative. The fusion of these two styles, positional play being confirmed by a crowing combination, breeds very effective results.[87]

---

[84] Adelson-Velskii, G., et all, Machine Plays the Game, Moscow: Nauka, 1983, pp. 84-85.

[85] Lasker, E., Lasker's Chess Manual, Moscow: OGIZ, Fizkultura i Turizm, 1937, p. 251.

[86] Lasker, E., Lasker's Manual of Chess, Philadelphia: Mc Kay Co, 1947, p.197.

To begin with, we distinguish two types of goal-directed positional problems: tactical and strategic. The first can be described in the following way: a goal is given over a limited set of material and positional parameters, it is remote, rather fuzzy and has a strong impact on the future (i.e., problem number 11). Preparing an attack against the opponent's king serves as an example of this kind of problem. A sequence of moves linking the goal with the current state is not only long but also fuzzy which forces us to set sublocal problems, i.e. intermediate local problems. Unlike combinational problems, intermediate local problems aim not at gaining material advantage but at improving current position by strengthening positional parameters. Sublocal problems can in turn be classified as key or non-key problems.

Using our previous classification of objective functions, the primary characteristic of a key sublocal problem is its strong impact on the rest of the game. Its other feature is that the objective function corresponding to a position which is not too remote from the current one is represented in the following form: positions of one or several "mobile" pieces is set as a goal. These pieces determine the feasibility of a successful attack against enemy king; the rest of the pieces remain stationed at their previous positions.

A famous game between T. A. Romanovskij and V. D. Ragozin played in the Second International Tournament held in Moscow in 1935 is an example of this kind of a problem.[88] Romanovskij asked Capablanca to comment on the game after it had reached a certain stage in the middle-game. Capablanca answered that the game can be won in "three moves". By a move he meant a key sublocal problem having its own objective function based on a number of constants representing positional parameters, a problem whose final goal is not too remote from a given position, and a problem solution of which has a strong impact on the future flow of the game (13 problem).

The objective of non-key sublocal problems is to improve one's position by "consciously" strengthening positional parameters. These problems are well defined and are evolutionary in nature (problem number 14).

But suppose we are unable to formulate a strategic positional problem. As E.

---

[87] This situation echoes the development of science which represents positional play and engineering which is combinational. Their evolution and mutual interaction has lead to the greatest advancement of scientific-technological progress.

[88] See Adelson-Velskii, G., et all, Machine Plays the Game, Moscow: Nauka, 1983.

Lasker wrote in his chess manual, when playing the positional style of chess it is extremely important to try to find out "what plan to follow in a balanced position when nothing provides a clear target of an attack or a defense?"[89]

Lasker points out that in this case it is necessary to accumulate small positional advantages as recommended by Steinitz. In other words, we have to set up local problems similar to the ones mentioned above - non-key sublocal goal-directed problems.

Note that in the event this kind of a problem defies formulation we have to switch to a protractive type of local problem.

§ 4. Com-positional problems

Two types of search problems looked at so far - goal and non-goal directed ones should not really be opposed to each other. In fact, they can interact in many interesting ways. One special case is a combinational type of a goal directed problem being used for material parameters and a non-goal directed problem for positional parameters. It was precisely this approach to local search problems that was proposed by M. Botvinnik.

> "In chess the goal of the inexact game is to win material. A similar goal may be found in an arbitrary game that models a control system, and in an arbitrary inexact task. To attempt to solve an inexact problem without having formulated the goal of the corresponding inexact game is to waste time. This goal is the basis of a strong algorithm for the solution of an inexact problem, and the basis for development of a deep and narrow tree. We shall see later why this is so.
>
> The goal of a game says what our aim is; only when we know this can we identify courses of action that cannot lead to our target, and exclude them from the tree. Knowing the goal lets us define the lines along which the search is to occur.
>
> The goal lets us direct the search; the scoring function lets us evaluate and stop a variation. The goal lets us form a search tree; the scoring function lets us strike a balance.
>
> The scoring function acts together with the goal of an inexact game and is therefore itself inexact. As distinct from the goal, which must be unique, the scoring function consists of two components: the first component allows us to evaluate the results obtained with the limits of the truncated tree; the second forecasts the possibility of reaching the goal beyond those limits.

---

[89] Lasker, E., Lasker's Chess Manual, Moscow: OGIZ, Fizkultura i Turizm, 1937, p.223.

The first component yields an exact answer (with the limited precision already determined) to the questions concerning the goals that have been reached; the second (the positional estimate)[90] gives a preliminary answer to the question of what will happen later, when the boundary of the truncation has moved further away. Taken together, these two components determine the value of a concluded variation."[91]

It seems to me that the com-positional type of problems is also typical of human behavior. Reducing them to a combinational type of problem represents an attempt to simplify one's behavior. Here, I would like to make another digression to discuss the issue.

Digression 2

It was during the sixties that my wife, my two sons - 15 and 8 years old, and I were coming back to Moscow after a vacation. We had to take a night train and the ride was about 12 hours long. Naturally, we wanted first class tickets. This way we would get private compartment with beds. In addition, tickets are not as expensive as they are for luxury class seats. But it so happened that we did not reserve our return trip tickets ahead of time; we did not really plan to go to that vacation spot, it turned out to be much less than perfect and so we decided to come back home and spend the rest of our vacation there.

When we arrived at the station I went over to buy tickets. The cashier told me that the station superintendent is in charge of selling first class tickets. The superintendent turned out to be a rather nice lady. After hearing my story she told me that she had only three tickets available in the sleeping car. One ticket would have to be in the regular car. I agreed. She gave me a permission to buy the tickets so I thanked her and went over and bought the tickets. I was touched by the woman's courtesy and wanted to to thank her. So I walked over to the station cafeteria to buy her some chocolates. As usual, there was a long line of people waiting to be served. Luckily, my two sons were there already. When they asked me what I was doing there I told them my intentions. They were quite shocked. "It is dumb, they said, to spend money on chocolates for some woman you don't really know who is really not expecting any gifts, whom we shall never see again and who could not do anything else for us anyway."

Well, I did not succumb to their rational arguments. I went ahead and bought the chocolates and gave them to the woman. She was very touched since she was not expecting

---

[90] According to Botvinnik, M., Solution of Inexact Problems, Moscow: Soviet Radio, 1979, positional value represents a ratio K_W/K_b where K_W and K_b is the number of squares in the trajectory controlled by white and black respectively. (p. 144).

[91] Botvinnik, M., Computers in Chess: Solving Inexact Search Problems, Springer-Verlag: New York, 1984, pp. 16-17.

any gifts at all. She asked me to wait a minute and left. When she returned she told me that she managed to find four tickets in one car. I thanked her sincerely and walked over to the ticket window to exchange the tickets. My children accompanied me. After looking at my old tickets, the cashier told me that I made a mistake by buying all adult tickets. She gave me new tickets plus some money back. The amount was much bigger than the price of the chocolates. My children were startled to see courtesy rewarded so quickly.

What is the point of the story? It seems I was solving a combinational problem having a clearly formulated goal: to get sleeping car tickets. The rules for acquiring the tickets were well defined: cashier - superintendent - cashier. The criterion of optimality was very clear: to get the tickets in a minimum amount of time. The problem could be solved by going through all these steps with well-defined material goals and choosing a ticket window with the shortest line. This is how my children were thinking, at least on the semi-intuitive level.

As for me, I took both relational parameters and material expenditures required for their realization into account. The weight of these parameters was incorporated into the final state along with the values of such material parameters as tickets. Here, the relational parameter of special importance was mutuality; it was her emotional reaction as such (actually, this applies to all positional parameters) that was important to me independently of the final material result.

Since the weights of positional parameters and material expenditures at the final and intermediate stages were unknown to me, the problem of ticket acquisition (that stage of it anyway) transformed into a non-goal directed problem with the values of parameters unknown.

The problem acquired com-positional nature. Weights assigned to positional parameters as well as to material expenditures were based on my personality. Another person with a different personality and a different system of values might have acted differently under the same circumstances and his behavior would have also been totally justified. There is no reason to think that everyone must assign importance to such parameter as mutuality. The reward received for this kind of behavior was largely a result of a series of happy coincidences (the lady in charge happened to have some extra tickets, she was sensitive to another person's thoughtfulness, etc). It does not follow that the reward had to come immediately. In fact, it could have never come at all! Do not discussions concerning righteous behavior revolve around the same issue? I think that the existence of individuals with different systems of values is very important for development. Therefore, acting under the conditions of uncertainty and "trying to be yourself" (trying to preserve one's individuality) represents a "continuous" growth of potential of the individual in question. Variety makes total potential of people as a whole grow as well.

Now, what about the solution of the problem? It reduces to some well known methods of solving similar problems under the conditions of uncertainty. Solution proceeds on contingency planning basis: at each step we try to choose the best possible course of action based on the current situation.

§ 5. Stages of the game

Our next topic deals with the stages of the game. Each stage can be viewed as a macrolocal problem(just as in biology, a certain set of molecules produces a micelle). Once again, we follow the systems approach in selecting the criterion for stage classification.

We shall consider the functional aspect of the problem first; this approach is player rather than machine oriented.

A well known grandmaster D. Bronstein proposed a very interesting classification of game stages based on the functional approach. These are the stages of the game as perceived by each side: 1) creation of a shelter for the king - a well protected command post; it is neither a castle nor is it a pill box since "a king is not going to fire any shells from this fortress of his "; 2) formation of pawn chains - the skeleton of the position; 3) "allocation" of the pieces, one's mobile resources according to their "personalities" and the intentions of the player; 4) crossing the "equator" for the purpose of invading enemy territory and destroying or weakening his position; 5) "active defense" is when an opponent attacks a given player and it becomes necessary to weaken the former to a degree when "he admits his defeat after the smoke has cleared "; 6) attack against the enemy king - this involves breaking through the pawn chains, capturing space, attacking weak pawns and advancing one's own pawns with support from the pieces.

D. Bronstein proposed a number of techniques to help fulfill the above mentioned functions. For instance, "creation of a shelter for the king" can be implemented not just by placing a king near the edge of the board (as we know, at the start of the game the king is placed in the center). Bronstein recommends not just building a "shelter, but a castle

fortified with double walls made of pawns. A king should be provided with personal bodyguards - a knight, a bishop, a rook, and four pawns."[92]

Bronstein proposes to build such a structure by means of a series of moves which includes castling. (p. 18).

Bronstein's approach seems original but lack of literature discussing his ideas makes it impossible for me, an amateur in the game, to judge the validity of his approach.

I want to focus on a more traditional method of stage classification. It is based on the structural features of a position. This approach does not ignore the functional aspect of the game but it emphasize structural characteristics. I think it makes sense to combine the functional and the structural approaches in order to gain deeper understanding of the overall problem.

In my opinion the structural approach incorporates two parameters: total number of pieces on the board and their mobility (which reflects both the position of one's own pieces as well as the opponent's). Here, it is assumed that each of these parameters is represented as a dichotomy, i.e. it can only assume two values. Matrix-table 8 shows all possible combinations of these two structural dimensions and their respective values.

Table 8. Mobility and number of pieces

Mobility	Number of pieces	
	Large	Small
Mobile	middle game	final endgame
Immobile	opening	early endgame

These four combinations correspond to traditional classification of the stages of the game, i.e. the opening, the middle game and the endgame;[93] "early endgame" stage owes its inclusion to the fact that the accepted meaning of the term denotes "final endgame" stage.

---

[92] Bronstein, D., Teach Yourself the Game, Moscow: Fizcultura i Sport, 1981, p. 17.

[93] Simplifying things a bit, we can say that human behavior strives to maximize the difference between positive and negative feelings.
Individual choose as to the course of action in a given environment is limited by his intellectual and emotional abilities. Using chess as an analogy, we can mention several different mechanisms used by the people of different ages in making a decision. These mechanisms are rather specific in the childhood stage and during the old age which is analogous to the opening and the endgame. It seems that the years between these two stages involve the use of more complex mechanisms. Here, we probably have some substages with their own peculiar features.

Here, I am going to skip over the processional aspect of classification of game stages.

In Table 9, I made an attempt to combine the structural and functional aspects of various stages with its process counterpart. Some ideas in this table are borrowed from the literature and some are my own.

Table 9. A multi-dimensional systems approach to the classification of the stages of a chess game

Stage of the game	approach underlying the definitions		
	functional	structural	process
Opening	1. Development of the pieces* 2. Mobilization of the pieces**	1. Limited mobility for a large number of pieces	1. Deep analysis of a large number of variations; knowing them allows both players to maintain equilibrium up to a certain point.
Middle game	Trying to achieve positional superiority***	1. A large number of mobile pieces 2. Passive kings****	Determining the best combination between the depth and the width of the search to ensure an optimal solution
Endgame	The goal can be set up as a terminal position, i.e. a mating position (or clearly won position)	1. A small number of mobile pieces**** 2. Kings are active in the struggle *****	A position whose outcome - a win or draw, can be verified by an analysis with both sides playing their best moves******

*Lasker, E., Lasker's Chess Manual, Moscow: OGIZ, Fizkultura I Turizm,1937.
**Capablanca, H. R., Chess Manual, Moscow: Fizkultura I Sport, 1975, p.31.
***Ibid, p. 36.
****Ibid, p.42.
*****Lisitzyn, G., The Closing Stage of the Game, Leningrad: Lenizdat, 1956, p. 5.
******Portish, L., 600 Endings, Moscow: Fizkultura I Sport, 1979, p.6

Although the above table is largely based on the structural features of a position, its multi-lateral approach puts the peculiarities of each stage in relief. For example, the structural connection between a large number of rather immobile pieces and a method of play based primarily on prior experience (rather than optimization or self-learning) allows us to define the opening stage of the game more precisely. Structural characteristics of this stage could, as such, apply to other stages as well. There exist middle- game positions having a large number of rather immobile pieces. The connection between the structure and the process based on previous experience links these characteristics with the function of piece mobilization thus giving a holistic definition of the opening stage.

We should also note that the structural aspect of the game does not determine either the function or the process: it merely brings greater degree of correlation between them.

The endgame stage is characterized by a small number of pieces on the board. This obviously simplifies the task of predicting the "final position" since the number of possible variations is not so large.

A final position, i.e a position which leads to an immediate mate shall be termed terminal position.[94] The presence of terminal positions allows for the implementation of backtracking algorithms which start at the final position and work back to the initial ones. Here, the actual terminal position may have a direct impact on the criterion associated with each intermediate position. Since the number of terminal positions in chess is enormous, an algorithm is basically forced to move "forward" from the current to the final (relative to a given task) position. This gives rise to a very difficult problem of formulating a local criterion to evaluate intermediate positions.

These features of endgame stage reveal why this stage can be handled by an optimization algorithm[95] which ensures that a terminal position is reached in a minimum number of moves.[96]

Nevertheless, terminal positions can be set up even at the beginning of the game when there is a large number of immobile pieces. Example of this is the so called "Fool's mate": white to move and mate in three. In reality these considerations apply only to the beginners. At more advanced levels of play, terminal positions are so remote from the initial one that they defy clear-cut formulation not only in the opening but also in the middlegame.

---

[94] Terminal positions can occur at an earlier stage - a state which leads inevitably to a mating attack. This feature is incorporated into some computer chess programs which warn the opponent of a mate in a certain number of moves.

[95] "...while in the middlegame players' plans are largely determined by their tastes and fantasy, in the endgame they are determined bu the position in question. Everyone must more or less follow the same plan which is typical for a given position independently of his style or tastes. In fact, many endgame positions are nothing else than logical problems sometimes having a unique solution." (Averbasch, U., Things to Know About the Endgame, Moscow,1979).

[96] This may affect the "fifty move" rule which states that if no exchange of pieces takes place or no pawns are advanced then one side can claim a draw.

Chapter 3. Evaluation of a state: parameters of integration.

§1. Analytical representation of the holistic effect.

Evaluation of the current state must reflect the state's overall impact on the game. But does the holistic effect yield to an analytical representation, i.e. can it be expressed by means of independent evaluations of different parts of a position? Proceeding in a straight forward manner we would have to assume that all parts are homogeneous and have the same impact on the environment, i.e. they have the same weight (importance, value). Under these assumptions individual parts can simply be added together (two apples plus three apples equals five apples). The value arrived at represents one of the essential preconditions for further development. But suppose different parts are qualitatively different or, if they are the same, have different "weights." How then do we integrate them into a single whole? This can be accomplished if we take into account the "weights" of the parts (two apples weighing 80 grams each plus three apples weighing 100 grams each together weigh 460 grams; this characteristic is essential for further development). Determining (discovering) these weights is a rather difficult task. Physics has had the most success in this area. The situation with biological systems where these weights are represented by, say emotions, with economic systems where they assume the form of prices, and with a number of other fields is more complicated.[55]

So it is possible to express the holistic value of a given state through its parts analytically provided their integration proceeds by calculating their respective weights which reflect the marginal contribution of each part to the development of the whole. This is the way many mathematical procedures for solving optimization problems are structured. Condensed, global evaluations of different parts, namely the Lagrange multipliers generated by these procedures, are precisely these kinds of weights. They are global in a sense of allowing us to measure the marginal impact on the overall development of the system produced by independently varying one single variable at each local step. (Since chess pieces are discrete in nature and are rather small in number the idea of varying one single variable must be reconsidered). The global nature of Lagrange multipliers is due to the fact that the process by which they are formed takes all the constraints of the problem into account: the initial state, the final goal, and the rules of the "game".

The sum of the products of each component with its respective global value represents the level of system's development at a given stage; it measures the impact of the current state on the solution of the global problem.

If we look at chess from this perspective we can say that it is possible to give a global evaluation of a position in analytical terms at each local stage provided we know an exact algorithm of the game. This evaluation would reflect the impact of a given position on the outcome of the game. But what if such an algorithm is not available, what if there is no program linking parts with a whole in a rigid and precise fashion, how then do we determine the holistic effect of a given position?

May be we can try to apply the ideas of exact algorithms to determine approximate values of parameters used in evaluating the current state of the system. These values should be a good approximation of the weights generated by exact algorithms. Of course, it should be remembered that the problem of evaluating the overall position is the cornerstone of our analysis. Evaluation of position's individual parts or of a single piece remains an important issue as far as varying their values independently of each other, but it is essential to keep in mind the interrelationships between the parts and the whole.

What does all this mean? It means that we have to proceed from the general to the particular, i.e. from an overall position to individual pieces.

Two questions arise immediately: 1) how do we analyze the structure of a position, i.e. will it be sufficient to limit our information to such primary objects as pieces - material (together with their coordinates) or will we need additional primary objects? and 2) how do we measure the values of the pieces when we do not know their total worth relative to some final goal?

Affirmative response to the first question gives rise to two new questions which reveal the need to introduce new primary parameters to be used in the evaluation of a position. The two questions are 3) what is the role of this new set of primary parameters and 4) how to evaluate these new parameters so that they are commensurable not only among themselves but also with the weights of material objects which together provides us with a evaluation of the position as a whole?

Let us proceed to answer these questions.

The first question asks whether the information pertaining to such primary objects as material parameters is sufficient to evaluate a position. The answer to this question bears directly on the second (as well as third and fourth) question which deals with evaluation of the material in inexact algorithms. I want to repeat that the structure of inexact algorithms rules out the possibility of linking the current state with the end of the game, i.e. of deducing the overall values of the pieces which would be sufficient to evaluate a position. We are thus forced to take a different route: first we assign unconditional values to the pieces, i.e. values corresponding to some ideal situation. We then try to introduce additional parameters which are supposed to reflect the current situation on the board, its specificity. This really represents an attempt to make up for the incompleteness of evaluations by including unconditional weights of the pieces. Now evaluation of these additional parameters which could be called compensatory parameters must account for their impact on the future flow of the game. Here parameters used to compensate for this deficiency pertain not only to individual pieces but also to groups of pieces or to the position as a whole. It seems to me that the nature of these parameters lies in the interrelations among the pieces in a given situation. It is precisely these relations which were not incorporated in the unconditional weights of the pieces. The more complete this set of parameters and the better their weight assignment the closer (other conditions being equal) is our evaluation of a position to that of an exact algorithm.

---

[55] The difficulty of determining the exact relative weights of all the parameters lead to an attempt to simplify the problem in the following way: to try to discover one dominant part and use it to evaluate the impact of the entire state on the future. (In sciences, this idea is known as using one single characteristic to determine the overall state). But such dominant characteristics are rather rare.

Therefore lack of information on the total value of the pieces (this information is needed to evaluate a position) can be compensated by determining the value of the entire position by means of unconditional values assigned to the pieces coupled with the parameters characterizing the relations among the pieces.

The next section explores the first component of the problem.

§ 2. Evaluation of material

Our preliminary discussion will focus on the analytical construction of unconditional values of the material. One reason for this is the availability of literature devoted to this topic. Another reason is the uniqueness, as I see it, of chess in its approach to determining the weight of the material.

Traditional treaties on chess - those devoted to the theory and practice of the game, are not concerned with the analytical methods of calculating the unconditional values of the pieces. For instance, E. Lasker's now classical chess manual does give us the unconditional weights of the chess pieces. The ratios presented in the above text are well known to all chess players. The author points out that these values are valid only in equal positions. The ratios between the values of the pieces can vary considerably depending on

the specific situation on the board. Just like many other players Lasker believes practical experience to be the source of the weights assigned to different pieces. The following quote from Lasker's book not only confirms my statement but also presents some possible relative numerical values of the pieces. These values will be important to us in the future.

> "...the masters, and among them certain regularities show very plainly. The experience derived from games played by such as deserve the title of masters, during centuries has proven these regularities."[56]

In his concluding remarks which really stand apart from the rest of the book Lasker gives us an interesting comparison between the values arrived at on the basis of practical experience and those produced by the mathematical analysis of the game.[57] He points out that Euler who was a great mathematician but a rather poor chess player was able to determine the values of the pieces mathematically and these values happened to correspond closely to the values representing the practical experience of great players.

The procedure used by Euler to determine the values of the pieces is not known. But Lasker points out that Euler's method "bears the seed of future discoveries"[58] Lasker does not give reference to Euler's work which could have been used to reconstruct his chain of reasoning using our present-day knowledge of the subject.[59]

Judging by Lasker's comments it is reasonable to assume that Euler used rather static positions to compute the values of the pieces, i.e. a case where material plays a decisive role. Two beginners playing in a random like fashion start from a given position and play n-games (n approaching infinity); material advantage for either side is determined by the winner of the match.

Lasker also notes that for Euler the mobility of a piece was probably the dominant characteristic in establishing its material value. Mobility, in turn, is determined by the free space or by the "obstacles" (thus a knight, although it is less mobile than the bishop, has an advantage over the bishop in being able to jump over the "barriers" set up by the opponent).

Well-known attempts at elaborating analytical methods to calculate values of the pieces were undertaken in non-chess literature.

Results derived from these non-chess sources coincide with practical experience to a surprising extent. This analysis is very valuable since it reveals the reasons why pieces have different weights. The weakness of these results lies in the fact that practical experience probably incorporates other important factors which may cancel each other.

---

[56] Lasker, E., Lasker's Manual of Chess, Philadelphia: Mc Kay Co, 1947, p. 35.

[57] Ibid., pp. 340-341.

[58] Lasker, E., Lasker's Chess Manual, Moscow: OGIZ, Fizkultura i Turizm, 1937, p.323.

[59] My own attempts to find this work were unsuccessful. I even asked specialists on Euler from the East German Academy of Sciences but they were unable to help me either.

The groundwork in the field of unconditional evaluation of the pieces (which happened to coincide with practical experience) was an article by H. Taylor.[60] In this work (I shall quote from it without further reference) Taylor proposed two methods of measuring the relative material values of the pieces. Common to both of these methods (this goes for Euler's method as well) was the dominant role of the material. Taylor wanted to reduce his experiment to the simplest possible case: the only pieces on the board were a king for one side and one piece for the other (besides king and pawns). Each piece wants to check the opponent's king. This is the set up of the problem:

> "A king and a piece of different colors are placed at random on two squares of a chessboard of n^2 squares: it is required to find the chance that the king is in check."

Unlike Euler's experiment, this method of calculating the relative values of the pieces for a given size board is based on a finite number of trials since the number of squares that can be occupied by a king and one piece is limited by the size of the board.

Taylor discovered that his method of calculating relative weights of the pieces produces results different from those used by the players. So he modified his experiment. Instead of calculating the probabilities of king checks he introduced an additional requirement of a "safe check." "Safe check" differs from a "simple check" in that the piece making the check cannot be captured by the opponent's king. The notion of a "safe check" incorporates two aspects: the mobility of the piece plus its ability to "overcome obstacles." In order to determine the relative weights of the pieces which E. Lasker spoke about Taylor elaborated a formula to calculate the probability of a king being in check, both "safe" and "simple." Table 3 below is borrowed from Taylor's article. It shows the results of his experiment for any size board having an even number of squares and for a regular board having 64 squares. Here, n stands for the number of squares on the board.

Ordering (reducing to a common denominator) the probabilities of different pieces putting a king in check gives us the following sequence: for the case of a "simple check" the values are 3, 5, 8, and 13 for the knight, the bishop, the rook, and the queen respectively; for the case of a "safe check" they are 3, 3.25, 6, 9.25 (12, 13, 24, 37). The correspondence of the last set of figures to the ratios commonly used by the players (they are 3, 3, 5, 9) is amazing.

Table 3. The probability of a king being in check

	Chess board with n2 squares		Regular board	
	Normal	Safe	Normal	Safe
Knight	8(n-2)/n2(n+1)	8(n-2)/n2(n+1)	1/12	1/12
Bishop	2/3x2n-1/n (n+1)	2/3x(n-2)(2n-3)n (n+1)	5/36	12/144
Rook	2/n+1	2(n-2)/n(n+1)	2/9	1/6
Queen	2/3x5n-1/n(n+1)	2/3x(n-2)(5n-3)n (n+1)	13/36	37/144

Working within an even more general setting M. Kavesh suggested evaluating a pawn by the number of safe checks it can give from the eighth rank where it can be promoted to a queen. The results coincided with the established value of the pawn relative to other pieces.

Attempts independent of those by Euler and Taylor to determine the relative values of the pieces analytically were made in more recent times. The basic idea was to have a clear board with one single piece. Then an average number of moves that a piece can make depending on its position on the board was calculated. For instance, a rook can make 14 moves from any square on the board. Obviously, an average number of moves it can make is also 14 (14x64/64). A queen can make 21 moves from 28 squares of the board, 23 moves from 20 squares, 25 moves from 12 squares, and 27 moves from 4 squares. Altogether a queen can make 1456 moves (21x28+23x20+25x12+27x4). Its average mobility is 22.75 moves (1456/64). Similar calculations can be performed not only for the knight and the bishop but also for the pawn and the king. Thus, a knight can make an average of 5.25 moves ((2x4+3x8+4x20+6x16+8x16)/64), a bishop can make 8.75 moves ((7x14+9x10+11x6+13x2)/32). As an autonomous piece, a king can make 6.5625 moves on the average ((3x4+5x24+8x36)/64); a pawn can make an average of 2.5 moves ((2x10+3x32+4x6)/48). Here, it is assumed that a pawn cannot occupy the first rank or move backwards from the eighth rank and the possibility of its promotion upon it reaching the eighth rank is ignored.

Below we have a formula for calculating an average number of possible moves. They are denoted by P_n for all the pieces[61] and the size of the board which is arbitrary is denoted by n.

---

[61] Derivation of the formula for the bishop is rather complicated so Gik recommends a book by Okunev, L. J. entitled Combinatorial problems on the chess board. Moscow: ONTI, 1935.

$P_n(K)=4(n-1)(2n-1)/n^2$
$P_n(Q)=2(n-1)(5n-1)/3n$
$P_n(R)=2(n-1)$
$P_n(B)=2(n-1)(2n-1)/3n$
$P_n(Kn)=8(n-1)(n-2)/n^2$
$P_n(P)=(n-1)(3n-4)/(n-2)$

Taking pawn as a basic unit of measurement the relative values of the pieces are 2.1 for the knight, 3.5 for the bishop, 5.6 for the rook, and 9.1 for the queen. These ratios in the values of the pieces are very close to those worked out by Taylor who used the "simple check" criterion (values arrived at by Taylor must be divided by 1.428 to even them out with the ones generated by the above method). This fact is not a coincidence. Indeed, the formula for king checks differs from the formula for the average number of moves that can be made by a given piece in only one respect: the former method takes the average number of possible moves available to a given piece relative to the total number of squares on the board and the latter takes the sum total of all possible moves and divides it by the total number of squares on the board.

An example with a rook should help illustrate the point. Using Taylor's method (the "simple check" criterion), the number of checks that a rook can make is 14 no matter what the position of the king is, i.e. 2(n-1). On the average, a rook will be able to do so from (n^2-1) squares since one of the squares is always occupied by itself. Therefore, the probability of a rook check is $2(n-1)/(n^2-1)=2/(n+1)$. This formula differs from the one which calculates the number of possible moves $2(n^2-1)/(n+1)=2(n-1)$ by a multiple of (n^2-1). In other words, in calculating the overall probability of a check that a rook can make from a given square we discover that it is precisely 2(n-1). The same considerations apply to cases where the number of possible checks varies depending on the position of the piece. A reader can perform these exercises by putting (n^2-1) in the numerator of each of the formulas. A discrepancy between the value of (n^2-1) and 1.428 - a coefficient used to compare the values of the pieces calculated according to one of the aforementioned methods, arises as a result of the chosen standard (a particular unit to measure the relative values of the pieces).

Although we arrive at the same relative values using either the Taylor's formula for "simple checks" or the mobility of the piece formula, the assumptions inherent in these two approaches have very different implications. Taylor's formulation of the problem is more meaningful: it relates the value of the piece with the final objective of the game. This simplifies our task of finding the appropriate set of constraints to be incorporated into our experiment aimed at establishing the relative weights of the pieces and make our analytic results resemble those arrived at on the basis of practical experience. The other more abstract approach which is based on piece mobility does not guarantee affinity of the results. Still, it is advantageous in some respects; for instance, it allows both the king and the pawns to be evaluated as active pieces.

To sum up we can state the following: the relative values of the pieces are functional in that they reflect the potential contribution of a given piece to the final goal of the game. At the same time, these values are idealized, i.e. they are deduced under "perfect" conditions practically unattainable in the real game. These values are rather unusual being averaged out and completely unconditional. The choice of the averaged out unconditional values rather than the highest ones generated by the most favorable circumstances "softens" the unconditional nature of these values. It makes the game more concrete. For instance, it takes into account the possibility that in some distant future a king and one piece may come to be positioned on any squares of the board.

This approach to evaluating the pieces leads to conclusion that the value of a piece is always the same. To make its value more concrete, a piece has to be examined within the context of a specific position. This is where positional parameters come in. For instance, some players claim that the values of the pieces in the opening are different from the usual rating. This means that the values of the pieces change as a result of positional considerations. E. Lasker writes that in the opening the relative values of the pieces differ depending on their initial positions.[62] They also differ from the commonly accepted weights of the pieces. The second column of Table 4 gives the values of the pieces according to Lasker; other columns compare his values with the commonly accepted ones.

Unfortunately, Lasker does not discuss any analytical methods for generating the values of the pieces in the opening. Still, it is easy to see that one of the major factors responsible for making the opening values of the pieces different from their overall ones is the immobility of the pieces in the opening, the fact that some pieces limit the development of others.

It would be very interesting to do an experiment to represent the modified (Lasker's) values of the pieces in the opening analytically by introducing a number of positional parameters. I leave this exercise to the reader.

---

[62] Lasker, E., Lasker's Manual of Chess, Philadelphia: Mc Kay Co, 1947, pp.106-107.

Table 4. A comparison of the values of the pieces in the opening (suggested by Lasker) with the established values.

The piece	Its value in the opening	Averaged out value: opening, by type of piece	opening, rel. to a pawn	general	The difference between relative averaged out value in the opening and overall: absol.	%
Pawn: d or e	2	1.3	1	1	0	0
" c or f	1.5					
" b or g	5/4					
" a or h	1/2					
Knight	4.5	4.5	3.5	3	+0.5	+16.7
King bishop	5	4.75	3.65	3	+0.65	+21.6
Queen bishop	4.5					
King rook	7	6.5	5	4.5	+0.5	+11.1
Queen rook	6					
Queen	11	11	8.5	9	-0.5	-0.6

Concluding our discussion on evaluating material parameters, I want to repeat that our ability to deduce these values is largely due to the peculiar set up of the game discussed in the first part of the book. The only point to add to that discussion is that every piece in chess except a king can pose a direct threat to enemy king even if the piece itself is "not in danger," i.e. every object in chess has the capacity to be directly involved in achieving the final goal. This applies to pawns as well for they can be promoted to stronger pieces enabling them to fulfill the above mentioned function.[63]

§ 3. The structure of positional parameters

I hope the above analysis of the methods of forming unconditional material values has made it clear that other parameters have to be introduced in order to improve upon our unconditional values in each specific situation. Doing so will enable us to give an overall evaluation of the current position (or some possible future position). These additional parameters must reflect interrelations among the pieces. As I have already mentioned above the purpose of introducing these parameters is to do explicitly what an exact algorithm of the game does implicitly.

I want to start our discussion with primary positional parameters, i.e. parameters that cannot be decomposed any further with the tools presently available. Primary parameters can be grouped together to form aggregate positional parameters such as the

---

[63] All this leads to a very interesting question which still remains open: "To what extent can the methods used in chess to form unconditional weights of the pieces be used as a general principle in other systems having some universal intermediate objects, i.e. objects not directly related to the final goal even if such a goal does exist?"

strength of the center, the development of the right or the left flanks, and so on. I shall not touch upon this kind of parameters in the present work. I only want to note that a lack of research on the methods of aggregating-disaggregating chess positions represents a gaping hole in the field of computer chess. There is no doubt that good chess players utilize these kinds of procedures to a great extent.

Beginning with Steinitz chess players have paid a lot of attention to primary positional parameters. Examples of these parameters are scattered throughout E. Lasker's chess manual. For instance, on page 202 he examines such positional parameters as double pawns, backward pawns, and isolated pawns. Lasker also mentioned a number of positional parameters that have not received wide acceptance till this day. He stressed the importance of positional parameters usually associated with pawn structure in analyzing piece coordination. Elaborating upon Steinitz's ideas, Lasker points out

> "If, one the one hand, the distance of two pieces, for instance of two knights, is very great, a co-operation or a conflict between them is possible only in the future, unless they co-operate or contend with each other indirectly..."[64]

It seems that the first systematic list of positional parameters was put together by C. Shannon.[65] As far as I know, it is still the only systematic list of positional parameters to be found in the chess literature; therefore, it is given here in its entirety.

> "(2) Pawn formation:
> (a) Backward, isolated and doubled pawns.
> (b) Relative control of center (pawns at K4, Q4, B4).
> (c) Weakness of pawns near king (e.g. advanced KNP)
> (d) Pawns on opposite color squares from bishop.
> (e) Passed pawns.
> (3) Positions of pieces:
> (a) Advanced knight (at K5, Q5, B5, K6, Q6, B6), especially if protected by pawn and free from pawn attack.
> (b) Rook on open file, or semi-open file.
> (c) Rook on seventh rank.
> (d) Doubled rooks.
> (4) Commitments, attacks and options:

---

[64] Lasker, E., *Lasker's Manual of Chess*. Philadelphia: Mc Kay Co, 1947, p.239.

[65] Shannon, G., "Programming a Computer for Playing Chess", *The Philosophical Magazine*, vol.XLI, 1950.

(a) Pieces which are required from guarding functions and, therefore, committed and with limited mobility.

(b) Attacks on pieces which give one player an option of exchanging.

(c) Attacks on squares adjacent to king.

(d) Pins. We mean here immobilizing pins where the pinned piece is of value not greater than the pinning piece; for example, a knight pinned by a bishop.

(5) Mobility." (p. 274).

Authors of chess program "Caissa" came up with a more elaborate but still not systematized set of positional parameters. The first version of the program included 18 parameters.[66] Subsequently the list was increased to 31.[67] Classification of the parameters included in this list is rather simple. First of all, 12 parameters associated with squares and pawns were distinguished from those associated with squares and pieces. The authors had also come up with 10 parameters reflecting the dynamics of the game. Among these parameters were "attacking the opponent's strongest piece," "a light piece attacking a weak square in the opponent's position," etc.

This is what the authors wrote concerning the subject:

"...at first, a rook must attack a square lying on an open file, then it must occupy it, and finally "invade the gluttony row". The opponent, in turn, must try to prevent this from happening by creating a position devoid of such weaknesses."[68]

And finally they distinguished another very specific parameter "Bishop attacking any piece belonging to the opponent (except a pawn)" which is really a particular case of another parameter. A parameter called "Nail" which describes a particular piece and pawn formation was also introduced.

I want to propose another classification of positional parameters presented in the aforementioned works on the "Caissa" chess program. It is given in Table 5. I believe this classification is more systematic. Its matrix representation reveals the connections between the relations and the objects comprising them. Pawns, because they are rather special (for instance, they have less mobility than other pieces and form relatively static structures - the skeleton of the game), are allotted a special place in this classification.

---

[66] Adelson-Velskil, G., et all, "Programming the Game of Chess", Soviet Mathematical Proceedings, v. XXY, issue 2/152, 1970.

[67] Adelson-Velskil, G., et all, Machine Plays the Game, Moscow: Nauka, 1983.

[68] Ibi[illegible]

Table 5. Classification of positional parameters

Types of the relations	Object of the relations					
	Squares	Pawns	Squares and pawns	Pieces	Piece and pawns	Squares,Pieces and pawns
Configuration	12.Open file****	2. Phalanx 6. Isolated Pawn	3.Center pawn* 5. Passed pawn 9.Isolated pawn on a semi-open file*****	32. 'Nail'^^^^ Pin^^^^^	19. Light piece occupying opponent's weak square 20. Knight posted in the center 22. Queen (rook) on an open or a semi-open file 23. White rook on the seventh or eight rank++++	(2)(c) weakness of the pawns near kingΔ
Control	10. Weak squar**		4. Pawn attacking a center square	14. Scope + 15.Opponent's highest ranking piece under attack++ 16. Opponent's unprotected piece under attack+++17. Fork++++ 31. Bishop attacking any opponent's piece (except a pawn)	18. Light piece hitting opponent's weak square 21. Rook (queen) on an open or a semi-open file 24. A square next to the opponent's king under attack 25. Queen-king scope++++++ (4)(b) attacking pieces with possibility of exchange	
Quantity and variety	number of free squares					(3)(g) pair of rooks^^^ variety of pieces bishop&rook vs. knight&rook; queen&knight vs. queen&bishop; two bishops vs. two knights^^^^
Loosing opportunity to perform a special operation					27. King castled 28. King-side castle given up° 29. Queen-side castle given up°° 30. Castle given up°°°°	

Comments to Table 5.
Arabic numerals refer to the list of parameters taken from the book by Adelson-Velskii, G., et all,1983, Machine Plays the Game, Moscow: Nauka. If the Arabic numeral is accompanied by a letter index and both are in parenthesis, then they refer to the parameters taken from the 1950 work by C. Shannon. No number in front of the parameter means that it was borrowed from other sources referred to separately.
Upper indices indicate that the respective parameter is accompanied by a footnote; I introduced all these symbols only for the sake of convenience. The need for explanations arose because players and programmers sometimes use different terms for the same parameter and some positional parameters are expressed in chess jargon which may be unfamiliar to the reader.
Explanation of the parameters:
*For white - squares e4, e5,e6, d4, d5, d6; for black - squares e5, e4, e3, d5, d4, d3.
***(Hole) - a square under attack by an opponent's pawn and is not and cannot be defended by
ones own pawns unless it moves to a different file as a result of a capture.)" -
****(Weak pawn) - a pawn standing on a square that lies behind a "hole".
****Open file
*****File open to a given player - only opponent's pawn(s) on the file
+ "For each piece, including a pawn, the squares that come under its attack including those
occupied by its own pieces or pawns (squares it defends)."
++"A piece occupying a square that is under attack by an opponent's piece(s) but not defended
by one's own piece(s)."
+++"A piece occupying a square that is under attack by an opponent's piece(s) but not defended
by one's own piece(s)."
++++"(Fork) - opponent's piece attacking two undefended higher ranking pieces of a given
player."
+++++ Black rook on the first or siccing rank.
++++++ "A square that can be occupied by a queen if it was placed on a square currently
occupied by the king. Having occupied it, opponent's queen (from vertical and horizontal squares
also a rook, bishop from the diagonal squares) gives your king a check."
° A rook occupying the h1 square in the starting position has moved
°° A rook occupying the a1 square in the starting position has moved
°°° Both rooks or a king have moved
°°°° "Any piece or a pawn in front of the opponent's pawn occupying its original position
between the c and the f files. This structure splits the flanks of the opponent."
°°°°° A given player must protect the piece in question with another piece
^ For instance, an advanced king night pawn
^^ Here, they mean a "pin" where the value of the pinned piece does not exceed the pinner; for
instance, a bishop pinning a knight.
^^^ Variety of pieces, especially important in combinational play since it creates more
opportunities.
^^^^ According to Capablanca, a bishop and a rook are stronger than a knight and a rook, but a knight and a queen may turn out to be stronger than a bishop and a queen; two bishops are almost always stronger than two knights.[69]

I believe by now the reader has sufficient empirical material to understand the diversity of positional parameters.

---

[69] Capablanca, H-R., Chess Manual, Moscow: Fizkultura i Sport,1975, pp. 30-31.

I want to make one additional remark. It is very unclear whether the above set of positional parameters is exhaustive. First of all, the language of the chess players is rich but it lacks precision. Therefore, just the formalization of parameters in chess players' arsenal presents certain difficulties: even qualified chess programmers are sometimes unable to do this.[70]

Assuming these difficulties can be overcome, there are other more serious problems. We can look at objects from an infinite number of perspectives so the number of positional parameters also seems to be infinite. Choosing a subset of positional parameters relevant to a given problem still remains an unresolved issue.

It is close to impossible to construct a system from which we can deduce the entire set of positional parameters which would be sufficient to create an "invincible" algorithm of the game.

Unfortunately, the set of positional parameters currently used by computer programs is still based on the experience of good players who express it in an intuitive manner. I think that one of the major obstacles to designing more effective algorithms is a lack of formalized procedures for developing new positional parameters. These procedures lie at the foundation of second-level programs, i.e. programs that modify other programs directly responsible for generating the moves.

Finally, let us consider the problem of evaluating positional parameters since their share in the overall value of a position may vary. I know of no analytical works on this subject that would even resemble analogous works on the evaluation of material parameters. In fact, I am familiar with two works which contain intuitive evaluations of positional parameters worked out by expert players.[71]

Whereas the first work presents only intuitive evaluations of positional parameters, the second one elaborates a number of ideas on how to form them. In Slate's work the general idea of assigning values to positional parameters (as well as material ones) is founded upon the following assumption:

> "Basically, the function adds several easily computable factors together. The dominating term is material. The remaining terms are rules of thumb designed to give the program a vague positional sense. Rather then expressing some theoretically

---

[70] An interested reader can consult the book by Adelson-Velskii, G., 1983, pp. 59-60, which discusses this aspect in detail.

[71] Adelson-Velskii, G., et.all, Machine Plays the Game, Moscow: Nauka,1983. (The last book by the same authors which provided material on positional parameters does not contain this kind of evaluation; and Slate, D., "The Northwestern University Chess Program", Chess Skill in Man and Machine. New York: Springer Verlag, 1977.

207

rigorous chess principles, these factors merely improve somewhat over 'aimless' moves. The design motive for the evaluator appears to be a self-fulfilling prophecy: we knew that it was going to be primitive and would thus depend heavily on tree searching. To do such deep tree searching would require that the evaluator do its work quickly, and so it wouldn't have the time to do anything clever, and thus would have to be primitive."[72]

The work quoted above encouraged others to try to determine at least partially (with a large number of parameters assigned intuitive values) the values of a few positional parameters analytically.

> For instance, "Knights are evaluated and king and center tropism and development. Center tropism is a peculiar measure of the closeness of a knight to the center of the board and is computed as 6 minus twice the sum of the rank and file distances from the center of the board to the square the knight is on. The value is then multiplied by 1.6...
>
> King tropism, for knights, is computed differently from that of rooks. We compute 5 minus the sum of the rank and file distances between the knight and enemy king and multiplied the result by 1.2... Like bishops, development of knights means not being on the back rank, the penalty being 9.4 ." (p.97).

Apparently, the weights of a number of positional parameters were not formed analytically: they were based primarily on players' intuition and experience.

Reasons underlying this method of evaluating positional parameters are deep. Subjectivity of judgement is rooted in the indeterminate nature of the game, in the need to adjust the play to the abilities and knowledge of a given player. Analysis of material parameters has revealed that they arose as as result of the rules of the game itself and their weights were calculated in an objective fashion. In general, subjective evaluation of a state is due to the subjective set of chosen positional parameters and their subjective evaluation. These considerations regarding the subjectivity of positional parameters do not rule out the possibility of establishing some averaged out magnitudes based on the analysis of a large number of games by players or machines. Still, these averaged out values would be mere guidelines, i.e they will reveal at first glance parameters which must be taken into account and call attention to those parameters whose subjective evaluations differ substantially from their averaged out norm; reasons for the gaps would have to be explained in order to justify the use of these averaged out values.[73]

M. Botvinnik proposed a totally different approach to positional parameters.

---

[72] Slate,D., and Atkin,L., " The Northwestern University Chess Program", Chess Skill In Man and Machine. New York: Springer-Verlag, 1977, pp. 93-94.

[73] This question is investigated in detail in the forthcoming book "Indeterministic Economics" with regards to evaluating the state of the firm.

Acknowledging the need to evaluate both material and positional parameters, he notes that:

> "Such a method (method by which various averaged out weights of positional parameters are introduced A.K.) of deriving a positional estimate in chess is wrong. A positional factor that yields a positive result in one situation may yield a negative result in another. For instance, are doubled pawns good or bad? The answer depends on the situation. Sometimes doubled pawns are a suitable target for an attack, since one of them cannot support the other. But sometimes doubled pawns contribute to the control of a square through which important communication lines (trajectories of pieces) pass, and then the doubling is extremely useful. The same thing may be said of other factors entering into the positional component of Shannon's prototypical scoring function.
>
> In the chess program under consideration, a quite different decision was made about the positional component of the scoring function, basing it on the control of those squares making up the trajectories that enter the MM. The side controlling the larger number of squares has a positional preponderance."[74] (p.29).

Botvinnik elaborates on this point:

> "Control of fields does not mean control of the whole board, but control of only those fields that may be used in the impending play. Therefore one must strive for control of the field consisting of those trajectories in which the pieces can move, but have not moved yet.
>
> At the node in the search tree where we find ourselves at a given moment, we must unravel all those sheaves of trajectories which have not yet been developed and determine which player has control of the majority of the fields consisting of the trajectories not yet used in the play. This allows us to forecast the result of the play -- the result of a search which, in particular, had to be renounced at the terminal nodes of the variations for lack of resources."[75]

An interested reader can refer to Botvinnik's book quoted above for a more detailed discussion of the methods of assigning weights to positional parameters. I do not intend to compare the two approaches to the evaluation of a position proposed by M. Botvinnik and C. Shannon (or his followers). Again, I refer the reader to G. Adelson-Velskii's book which presents a comparative analysis of Botvinnik's approach.

Finally, evaluation of a state calls for an answer to the following question: "How do we compare the value of material parameters with the value of positional parameters?" The answer given in the literature is very intuitive. Essentially, it claims the the total value of positional parameters should be somewhere between that of a half-pawn to two pawns. Adelson-Velskii notes that the value of the smallest material unit - a half-pawn "must be high enough to exceed the total values of all positional parameters."[76] At the same time, they say "It would be very beneficial to think of positional parameters comparable to the material in value."[77] D. Slate, D. writes "The total of the nonmaterial terms does not usually exceed the value of about a pawn and a half."[78] And finally a quote from M. Botvinnik: "We may assume that a so-called positional sacrifice should not exceed two units of material (this will be made more precise in an experiment)."[79]

---

[74] Botvinnik, M., Solution of Inexact Problems. Moscow: Soviet Radio,1979, p. 29.

[75] Ibid., p. 38.

H. Berliner notes:

> "I differ from some leading practitioners in that I have always been an advocate of positional contributions that can some to more than the value of a Pawn. We have had situations where Hitech considered one side to have almost 3 pawns worth of positional compensation."[80]

This emphasis on increasing the relative weight of positional parameters stems from the experience of professional players. In some cases, they are willing to sacrifice material and not just a pawn but a whole piece in order to improve their position. But here we are jumping ahead of ourselves. I will have a chance to discuss positional sacrifice when I talk about the relationship between combinational and positional styles of play.

---

[76] Adelson-Velskii, G., et all, Machine Plays the Game, 1983, p. 69.

[77] Ibid.

[78] Slate,D., and Atkin,L., " The Northwestern University Chess Program", Chess Skill in Man and Machine, New York: Springer-Verlag, 1977, p. 95.

[79] Botvinnik, M.M., Computers in Chess: Solving Inexact Search Problems, Springer-Verlag: NewYork, 1984, p. 38.

[80] Berliner, H., "Some Innovations by Hitech", ICCA Journal, Vol. 10, No. 3, 1987, pp. 111-117.

Chapter 4. Class of local problems and stages of the game.

§ 1. Classification of search problems

Having better understood the evaluation of a position we can discuss different types of local search problems. Problem type depends primarily on the structure and the value of the final product associated with each local problem. What determines the constraints of the problem is the current position on the board.[81]

The structure of the final product can be represented in a variety of ways. Our classification is based on four characteristics.

1. How fixed are the values of independent input variables, i.e. some or all of the variables are given as constants or as unknowns; all this is based on the hypothesis of the existence of the true values of the variables incorporated into the objective function.

2. Remoteness (number of moves separating) of the final position from the current one: the final position (relative to a given problem) can be distant or close (even one half-move away) to the current one.

3. The extent to which the structure of the objective function is certain: the qualitative characteristics of its arguments may or may not be well defined.

Table 6. The structure of the objective function of the local problem.

Characteristics of the future positon		Remoteness of the future position from the current one			
The values of the variables determining how fixed is the value of the position	The degree of uncertainty of the structure of of the position	Far		Near	
		Impact on the development		Impact on the development	
		strong	weak	strong	weak
Constants	Clear				
	Fuzzy				
Unknown	Clear				
	Fuzzy				

---

[81] Of course, this is only true if we assume that the process is markovian. Actually, it may be worth while in formulating local problems to remember previous moves made in the game. Here, I am going to ignore these considerations.

4. The variability of the value of a given parameter as a function of the impact of a current position on the future flow of the game.

Table6 presents all possible combinations of these characteristics in amatrix form.

Table 7 gives names and abbreviations for the 16 types of local problems derived in Table 6.

Table7. Types of local problems and areas of their application

No. Area of application

1. Positional problem with a highly effective procedure for eliminating inferior moves
2. Positional problem; the procedure for eliminating inferior moves is not very developed
3. Positional problem given in aggregate variables with a highly effective procedure for eliminating inferior moves
4. Positional problem given in aggregate parameters; the procedure for eliminating inferior moves is not very developed
5. Sublocal problem of the general form with an unexpectedly strong outcome
6. Non-key sublocal problem of the general form (standard algorithm block)
7. Sublocal problem given in aggregate parameters with an unexpectedly strong result
8. Sublocal problem given in aggregate parameters with a weak result
9. Positional goal directed type of problem with sublocal problems
10. Goal directed type of problem with sublocal problems, result weak
11. Goal directed type of problem given in aggregate parameters with sublocal problems; produces drastic changes on the board
12. Goal directed type of problem given in aggregate parameters with sublocal problems; weak results
13. Combinational problem; the amount of material captured is decisive
14. Non-key sublocal problem
15. x
16. Combinational problem; material is won and positional parameters vary.

I want to introduce based on this classification some general definitions pertaining to local problems. If parameters associated with the final position are given (they are known values) this position shall be called a goal.

A problem based on this kind of a goal shall be called a goal-directed problem.

If the number of moves separating the final position from the current one is more than one the problem shall be called protractive; a one-step problem shall be called a condensed problem.

Suppose all the initial as well as final conditions of a goal-directed protractive problem are given. Then the problem reduces to finding an optimal trajectory of moving from one state to the other. This trajectory could, for example, fulfill the requirement of reaching the final state in a minimum amount of time. A function designed to implement this kind of a constraint shall be termed a criterion of optimality. In an extreme case when the choice of the trajectory does not make any difference a local optimization problem reduces to finding just one such trajectory which of course must not violate the rules governing the transformation of one state into another.

Unlike goal-directed local problems local problems that include parameters of the final state whose values are unknown shall be called non-goal-directed problems. Here, a function which evaluates a final position shall be called an weight function. Non-goal-directed problems can be either protractive or condensed.

In setting up a local problem we may have some idea of the value of some part of the final position in question. Our hypothesis does not have to be incorporated into the evaluation of the final position explicitly; it can be represented as a constraint assigned a constant value. These types of constraints represent objective constraints. They should not be confused with global constraints which also stay fixed but which pertain to the local problem in its global form; these constraints are uncontrollable, they are completely determined by previous development.

All the terms associated with the definition of a final position (goal, objective constraints, criterion of optimality, weight function) combine to form the concept of an objective function. It allows us to focus on the general structure of the solution of the local problem in case we are not interested in all its details.

§ 2. Non-goal directed protractive problems.

Let us begin our analysis with non-goal directed type of local problems.

Constraints incorporated into local non-goal directed condensed problems are determined by the current position of the pieces (both one's own and the opponent's). The value of a move is determined by the weight function as the difference between the evaluation of the new position (one half-move) by a given player and his evaluation of opponent's half-move. It is assumed that the opponent responds to your own best half-move with his best half-move. This method of evaluation is based on Zermelo's principle. It states that the value of a move is determined by the maximum value of a position reachable by that move.

All this is rather trivial. It is analogous to a firm trying to maximize its profit or a man trying to maximize the difference between his positive and negative emotions. (It is also akin to the process of setting up local problems in the above mentioned decompositional algorithms of optimization).

In constructing a weight function for a given problem it is assumed that the values of all material and positional parameters included in it are unknown, i.e. all its arguments are variables or unknown quantities. In other words, we cannot predict the exact values of all the parameters determining the value of the position at the end of the search; i.e. we do not know which pieces shall stay on the board, what squares they will occupy, or which positional parameters associated with that particular configuration will come into play.

At the same time, it is assumed that all the coefficients of the arguments, i.e. weights of material and positional parameters are completely determined. The other two components of the objective function cannot be determined outside a given situation. For instance, the "distance" separating the final position from the current one largely depends on the search algorithm being used, the ability to eliminate inferior moves, and the power of the computer. The impact of unknown structure on the future of the game is usually rather weak, i.e it is not meant to accommodate major changes. At the same time, it can only be determined post factum or after the problem has been solved.

If we knew the overall values of parameters included in the local objective function, i.e. values conveying global information about the entire game, then the solution of such a one-step local problem would constitute one segment of the optimal trajectory representing the solution of the global problem. But as I have mentioned many times before such global evaluations cannot be deduced within an algorithm of an inexact game: we have to settle for partial evaluations. Therefore the value of the local objective function is not exact. This has lead many people involved in the development of chess algorithms to expand the search horizon, i.e. increase the depth of the search. It is clear that the lesser the depth of the search the greater the probability, all other conditions being equal, of undiscovered flaws in the positions beyond the horizon: evaluation of a position at a given step serves merely as a guideline for future play.

Now, since the values of the variables comprising the objective function are not known, the local problem at the start of the procedure would, in theory, involve just one move - two half-moves. The actual number of moves looked at, i.e. the horizon of the protractive problem depends on several aspects including the power of the algorithm and the machine being used. It was precisely this uncertainty over the length of the sequence of local one-move problems that stimulated me to call this sequence of local problems a protractive local non-goal directed problem.

As I have noted in the section on general description of a chess algorithm, the number of possible moves in the game is enormous so the machine is really limited in its ability to exhaust all the variations. Therefore, the solution involves a well-known trade off between the depth of search or the horizon and the width of search or the number of moves examined at each step.

It was precisely this kind of approach to the formulation of a local problem and a corresponding objective function (viewed within the global framework of the entire algorithm) that was suggested by Shannon, C., 1950. This approach, somewhat modified, is still being used today by most algorithms of the game.

At the same time, there are strong arguments against Shannon's approach. M. Botvinnik thinks that

> "The first method, however, is hopeless as means of finding a good solution to an inexact task, as we shall show by applying it to the game of chess. On the average, each side in a chess game has 20 available moves. Suppose we start in a position where White is to move; all 20 moves of White's pieces must be included in the search tree. Then it is Black's turn; for each of White's 20 moves, Black has 20 answers, so that the tree contains 400 Black moves. If we extend the variation to include another move by White, the tree will complete moves -- six half-moves ("plies") -- it will contain some 67,000,000 nodes! The size of the tree is an exponential function of the depth.
>
> If we develop a suitable algorithm for a chess program, the resources of the computer (memory and speed) are less significant than might be supposed at first glance. Whenever we extend the depth of the variation by a move, the resources of the machine must increase by several factors of 10; this is impossible in practice. Thus the main element is not the power of the computer, but the skill of the programmer, i.e., the extent to which he has contrived to depart from the fundamental principle (a full-width search) in order to economize on machine resources."[82]

I agree with Botvinnik's criticism of protractive problems but I do not think they should be ignored altogether. Not only is the protractive approach capable of competing with other approaches in formulating local problems, but it can also be used in conjunction with other techniques. I shall elaborate on this idea further on in the chapter.

§ 3. Goal-directed protractive problems

In the present section, rather than concern myself with the ways of improving the set up and the solution of non-goal directed protractive problems, I shall try to apply other principles to resolve these problems, principles rooted in the goal-oriented approach especially its protractive variety.

In any goal-directed problem it is assumed that the goal can be reached without breaking the rules of the game. A reader may be surprised by this seemingly naive and indispensable requirement. But there is a branch of chess theory called retrograde chess.[83] There are positions that are legal as far as the positions of the pieces are concerned but which can never be reached from the starting position if we follow the rules of the game. The purpose of retrograde analysis is to prove that there exists no backward sequence of legal moves from a given position (legal in itself) that would eventually lead back to the starting position.

Therefore, when we form a long-term goal we have to be careful to avoid "retrograde goal".

Assume there is no risk of such a goal being set (it does not occur frequently anyway so we can ignore it for the time being). Still we have to set a goal that is realistic, not just wishful thinking on our part (that is random in the extreme case). So far, there is no formal procedure to accomplish this task. I shall attempt to show below that the solution to this problem calls for a rather sophisticated analysis of the structure of the state (current position) - uncovering positional features peculiar to it. This would ensure that the goal set is more or less reasonable.

We shall begin our analysis with attractor type of problems. Its impact on the outcome of the game is strong and its specific goal is given over a limited set of material parameters (problem number 13). A typical goal for this kind of problem is to try to capture opponent's material which would virtually defeat his position.

In other words, positional considerations can be ignored when the goal is to capture a lot of enemy material. The criterion of the local problem reduces to finding an optimal trajectory linking the current position with the final state to be attained.

It is customary to call this kind of problem combinational.

---

[83] Smullyan, R., The Chess Mysteries of Sherlock Holmes. Hutchinson, 1980. This branch of chess, sometimes called "chess logic" is also becoming part of computer chess. See Alden, B. and Bramer, M., "Development of a Problem for Solving Retrograde Analysis Chess Problems", Advances in Computer Chess, 3, ed. by M. R. B. Clarke, Oxford: Pergamon Press, 1982, pp. 121-137.

The section entitled "Forced play and how to prepare for it" of Adelson-Velskii's book[84] quoted many times presents some interesting examples of a goal (only in simple cases like mate in one) which consists of capturing enemy material without searching through all the variations branching from the current position.

It is easy to see that combinational problems have some very profound advantages: its solution gives one side an upper hand by generating a swift attack aimed at winning enemy material.

Nevertheless these advantages are rather conditional, they are realizable only under some very special circumstances. Although the goal associated with a combinational type problem is relatively clear, it suffers from a number of drawbacks which stem from the need to determine the values of different pieces. Of course, if one side is up a heavy piece then the evaluation of a position is generally not a problem. Here, "generally" excludes, among other things, an intentional sacrifice on the part of the opponent. (The subject of a sacrifice in chess will be addressed in the next chapter.)

But what happens if the consequences of a combination are unclear or if we cannot come up with such a combination at all?

> "Maybe a combination does not exist at all, may be he [player] is just chasing a ghost? A search for a combination may turn out to be futile. To base one's play on the existence of a combination is a very dangerous enterprise."[85]

How then is the game to be played? It must be played positionally. Positional play does not preclude combinations. On the contrary, it creates predisposition towards a combination. To quote Lasker once again

> "Hence, he [Steinitz] concluded that some characteristic, a quality of a given position must exist that to a discerning eye would indicate the success or the failure of the search before it was actually undertaken. And this characteristic, if explicable by reason, of what could it possibly consist if not an advantage or a disadvantage...And an advantage, if reasonable, what could that be except the thing that was generally termed so: greater material force, greater mobility, greater effect against the King..."[86]

So positional play makes sense if it eventually leads to a combination. To rely exclusively on combinations is flawed for it rings of primitive pragmatism. Purely positional style has the disadvantage of being over speculative. The fusion of these two styles, positional play being confirmed by a crowing combination, breeds very effective results.[87]

---

[84] Adelson-Velskii, G., et all, Machine Plays the Game, Moscow: Nauka, 1983, pp. 84-85.

[85] Lasker, E., Lasker's Chess Manual, Moscow: OGIZ, Fizkultura i Turizm, 1937, p. 251.

[86] Lasker, E., Lasker's Manual of Chess, Philadelphia: Mc Kay Co, 1947, p.197.

To begin with, we distinguish two types of goal-directed positional problems: *tactical* and *strategic*. The first can be described in the following way: a goal is given over a limited set of material and positional parameters, it is remote, rather fuzzy and has a strong impact on the future (i.e., problem number 11). Preparing an attack against the opponent's king serves as an example of this kind of problem. A sequence of moves linking the goal with the current state is not only long but also fuzzy which forces us to set *sublocal* problems, i.e. intermediate local problems. Unlike combinational problems, intermediate local problems aim not at gaining material advantage but at improving current position by strengthening positional parameters. Sublocal problems can in turn be classified as *key* or *non-key problems*.

Using our previous classification of objective functions, the primary characteristic of a key sublocal problem is its strong impact on the rest of the game. Its other feature is that the objective function corresponding to a position which is not too remote from the current one is represented in the following form: positions of one or several "mobile" pieces is set as a goal. These pieces determine the feasibility of a successful attack against enemy king; the rest of the pieces remain stationed at their previous positions.

A famous game between T. A. Romanovskij and V. D. Ragozin played in the Second International Tournament held in Moscow in 1935 is an example of this kind of a problem.[88] Romanovskij asked Capablanca to comment on the game after it had reached a certain stage in the middle-game. Capablanca answered that the game can be won in "three moves". By a move he meant a key sublocal problem having its own objective function based on a number of constants representing positional parameters, a problem whose final goal is not too remote from a given position, and a problem solution of which has a strong impact on the future flow of the game (13 problem).

The objective of non-key sublocal problems is to improve one's position by "consciously" strengthening positional parameters. These problems are well defined and are evolutionary in nature (problem number 14).

But suppose we are unable to formulate a strategic positional problem. As E.

---

[87] This situation echoes the development of science which represents positional play and engineering which is combinational. Their evolution and mutual interaction has lead to the greatest advancement of scientific-technological progress.

[88] See Adelson-Velskii, G., et all, *Machine Plays the Game*, Moscow: Nauka, 1983.

Lasker wrote in his chess manual, when playing the positional style of chess it is extremely important to try to find out "what plan to follow in a balanced position when nothing provides a clear target of an attack or a defense?"[89]

Lasker points out that in this case it is necessary to accumulate small positional advantages as recommended by Steinitz. In other words, we have to set up local problems similar to the ones mentioned above - non-key sublocal goal-directed problems.

Note that in the event this kind of a problem defies formulation we have to switch to a protractive type of local problem.

§ 4. Com-positional problems

Two types of search problems looked at so far - goal and non-goal directed ones should not really be opposed to each other. In fact, they can interact in many interesting ways. One special case is a combinational type of a goal directed problem being used for material parameters and a non-goal directed problem for positional parameters. It was precisely this approach to local search problems that was proposed by M. Botvinnik.

> "In chess the goal of the inexact game is to win material. A similar goal may be found in an arbitrary game that models a control system, and in an arbitrary inexact task. To attempt to solve an inexact problem without having formulated the goal of the corresponding inexact game is to waste time. This goal is the basis of a strong algorithm for the solution of an inexact problem, and the basis for development of a deep and narrow tree. We shall see later why this is so.
>
> The goal of a game says what our aim is; only when we know this can we identify courses of action that cannot lead to our target, and exclude them from the tree. Knowing the goal lets us define the lines along which the search is to occur.
>
> The goal lets us direct the search; the scoring function lets us evaluate and stop a variation. The goal lets us form a search tree; the scoring function lets us strike a balance.
>
> The scoring function acts together with the goal of an inexact game and is therefore itself inexact. As distinct from the goal, which must be unique, the scoring function consists of two components: the first component allows us to evaluate the results obtained with the limits of the truncated tree; the second forecasts the possibility of reaching the goal beyond those limits.

---

[89] Lasker, E., *Lasker's Chess Manual*, Moscow: OGIZ, Fizkultura i Turizm, 1937, p.223.

The first component yields an exact answer (with the limited precision already determined) to the questions concerning the goals that have been reached; the second (the positional estimate)[90] gives a preliminary answer to the question of what will happen later, when the boundary of the truncation has moved further away. Taken together, these two components determine the value of a concluded variation."[91]

It seems to me that the com-positional type of problems is also typical of human behavior. Reducing them to a combinational type of problem represents an attempt to simplify one's behavior. Here, I would like to make another digression to discuss the issue.

Digression 2

It was during the sixties that my wife, my two sons - 15 and 8 years old, and I were coming back to Moscow after a vacation. We had to take a night train and the ride was about 12 hours long. Naturally, we wanted first class tickets. This way we would get private compartment with beds. In addition, tickets are not as expensive as they are for luxury class seats. But it so happened that we did not reserve our return trip tickets ahead of time; we did not really plan to go to that vacation spot, it turned out to be much less than perfect and so we decided to come back home and spend the rest of our vacation there.

When we arrived at the station I went over to buy tickets. The cashier told me that the station superintendent is in charge of selling first class tickets. The superintendent turned out to be a rather nice lady. After hearing my story she told me that she had only three tickets available in the sleeping car. One ticket would have to be in the regular car. I agreed. She gave me a permission to buy the tickets so I thanked her and went over and bought the tickets. I was touched by the woman's courtesy and wanted to to thank her. So I walked over to the station cafeteria to buy her some chocolates. As usual, there was a long line of people waiting to be served. Luckily, my two sons were there already. When they asked me what I was doing there I told them my intentions. They were quite shocked. "It is dumb, they said, to spend money on chocolates for some woman you don't really know who is really not expecting any gifts, whom we shall never see again and who could not do anything else for us anyway."

Well, I did not succumb to their rational arguments. I went ahead and bought the chocolates and gave them to the woman. She was very touched since she was not expecting any gifts at all. She asked me to wait a minute and left. When she returned she told me that she managed to find four tickets in one car. I thanked her sincerely and walked over to the ticket window to exchange the tickets. My children accompanied me. After looking at my old tickets, the cashier told me that I made a mistake by buying all adult tickets. She gave me new tickets plus some money back. The amount was much bigger than the price of the chocolates. My children were startled to see courtesy rewarded so quickly.

What is the point of the story? It seems I was solving a combinational problem having a clearly formulated goal: to get sleeping car tickets. The rules for acquiring the tickets were well defined: cashier - superintendent - cashier. The criterion of optimality was very clear: to get the tickets in a minimum amount of time. The problem could be solved by going through all these steps with well-defined material goals and choosing a ticket window with the shortest line. This is how my children were thinking, at least on the semi-intuitive level.

As for me, I took both relational parameters and material expenditures required for their realization into account. The weight of these parameters was incorporated into the final state along with the values of such material parameters as tickets. Here, the relational parameter of special importance was mutuality; it was her emotional reaction as such (actually, this applies to all positional parameters) that was important to me independently of the final material result.

Since the weights of positional parameters and material expenditures at the final and intermediate stages were unknown to me, the problem of ticket acquisition (that stage of it anyway) transformed into a non-goal directed problem with the values of parameters unknown.

The problem acquired com-positional nature. Weights assigned to positional parameters as well as to material expenditures were based on my personality. Another person with a different personality and a different system of values might have acted differently under the same circumstances and his behavior would have also been totally justified. There is no reason to think that everyone must assign importance to such parameter as mutuality. The reward received for this kind of behavior was largely a result of a series of happy coincidences (the lady in charge happened to have some extra tickets, she was sensitive to another person's thoughtfulness, etc). It does not follow that the reward had to come immediately. In fact, it could have never come at all! Do not discussions concerning righteous behavior revolve around the same issue? I think that the existence of individuals with different systems of values is very important for development. Therefore, acting under the conditions of uncertainty and "trying to be yourself" (trying to preserve one's individuality) represents a "continuous" growth of potential of the individual in question. Variety makes total potential of people as a whole grow as well.

---

[90] According to Botvinnik, M., Solution of Inexact Problems, Moscow: Soviet Radio, 1979, positional value represents a ratio K_w/K_b where K_w and K_b is the number of squares in the trajectory controlled by white and black respectively. (p. 144).

[91] Botvinnik, M., Computers in Chess: Solving Inexact Search Problems, Springer-Verlag: New York, 1984, pp. 16-17.

Now, what about the solution of the problem? It reduces to some well known methods of solving similar problems under the conditions of uncertainty. Solution proceeds on contingency planning basis: at each step we try to choose the best possible course of action based on the current situation.

§ 5. Stages of the game

Our next topic deals with the stages of the game. Each stage can be viewed as a macrolocal problem(just as in biology, a certain set of molecules produces a micelle). Once again, we follow the systems approach in selecting the criterion for stage classification.

We shall consider the functional aspect of the problem first; this approach is player rather than machine oriented.

A well known grandmaster D. Bronstein proposed a very interesting classification of game stages based on the functional approach. These are the stages of the game as perceived by each side: 1) creation of a shelter for the king - a well protected command post; it is neither a castle nor is it a pill box since "a king is not going to fire any shells from this fortress of his "; 2) formation of pawn chains - the skeleton of the position; 3) "allocation" of the pieces, one's mobile resources according to their "personalities" and the intentions of the player; 4) crossing the "equator" for the purpose of invading enemy territory and destroying or weakening his position; 5) "active defense" is when an opponent attacks a given player and it becomes necessary to weaken the former to a degree when "he admits his defeat after the smoke has cleared "; 6) attack against the enemy king - this involves breaking through the pawn chains, capturing space, attacking weak pawns and advancing one's own pawns with support from the pieces.

D. Bronstein proposed a number of techniques to help fulfill the above mentioned functions. For instance, "creation of a shelter for the king" can be implemented not just by placing a king near the edge of the board (as we know, at the start of the game the king is placed in the center). Bronstein recommends not just building a "shelter, but a castle fortified with double walls made of pawns. A king should be provided with personal bodyguards - a knight, a bishop, a rook, and four pawns."[92]

Bronstein proposes to build such a structure by means of a series of moves which includes castling. (p. 18).

Bronstein's approach seems original but lack of literature discussing his ideas makes it impossible for me, an amateur in the game, to judge the validity of his approach.

I want to focus on a more traditional method of stage classification. It is based on the structural features of a position. This approach does not ignore the functional aspect of the game but it emphasize structural characteristics. I think it makes sense to combine the functional and the structural approaches in order to gain deeper understanding of the overall problem.

In my opinion the structural approach incorporates two parameters: total number of pieces on the board and their mobility (which reflects both the position of one's own pieces as well as the opponent's). Here, it is assumed that each of these parameters is represented as a dichotomy, i.e. it can only assume two values. Matrix-table 8 shows all possible combinations of these two structural dimensions and their respective values.

Table 8. Mobility and number of pieces

Mobility	Number of pieces	
	Large	Small
Mobile	middle game	final endgame
Immobile	opening	early endgame

These four combinations correspond to traditional classification of the stages of the game, i.e. the opening, the middle game and the endgame;[93] "early endgame" stage owes its inclusion to the fact that the accepted meaning of the term denotes "final endgame" stage.

---

[92] Bronstein, D., Teach Yourself the Game. Moscow: Fizcultura i Sport, 1981, p. 17.

[93] Simplifying things a bit, we can say that human behavior strives to maximize the difference between positive and negative feelings.
Individual choose as to the course of action in a given environment is limited by his intellectual and emotional abilities. Using chess as an analogy, we can mention several different mechanisms used by the people of different ages in making a decision. These mechanisms are rather specific in the childhood stage and during the old age which is analogous to the opening and the endgame. It seems that the years between these two stages involve the use of more complex mechanisms. Here, we probably have some substages with their own peculiar features.

Here, I am going to skip over the processional aspect of classification of game stages.

In Table 9, I made an attempt to combine the structural and functional aspects of various stages with its process counterpart. Some ideas in this table are borrowed from the literature and some are my own.

Table 9. A multi-dimensional systems approach to the classification of the stages of a chess game

Stage of the game	approach underlying the definitions		
	functional	structural	process
Opening	1. Development of the pieces* 2. Mobilization of the pieces**	1. Limited mobility for a large number of pieces	1. Deep analysis of a large number of variations; knowing them allows both players to maintain equilibrium up to a certain point.
Middle game	Trying to achieve positional superiority***	1. A large number of mobile pieces 2. Passive kings****	Determining the best combination between the depth and the width of the search to ensure an optimal solution
Endgame	The goal can be set up as a terminal position, i.e. a mating position (or clearly won position)	1. A small number of mobile pieces**** 2. Kings are active in the struggle *****	A position whose outcome - a win or draw, can be verified by an analysis with both sides playing their best moves******

*Lasker, E., Lasker's Chess Manual, Moscow: OGIZ, Fizkultura i Turizm,1937.
**Capablanca, H. R., Chess Manual, Moscow: Fizkultura i Sport, 1975, p.31.
***Ibid, p. 36.
****Ibid, p. 42.
*****Lisitzyn, G., The Closing Stage of the Game, Leningrad: Lenizdat, 1956, p. 5.
******Portish, L., 600 Endings, Moscow: Fizkultura i Sport, 1979, p.6

Although the above table is largely based on the structural features of a position, its multi-lateral approach puts the peculiarities of each stage in relief. For example, the structural connection between a large number of rather immobile pieces and a method of play based primarily on prior experience (rather than optimization or self-learning) allows us to define the opening stage of the game more precisely. Structural characteristics of this stage could, as such, apply to other stages as well. There exist middle- game positions having a large number of rather immobile pieces. The connection between the structure and the process based on previous experience links these characteristics with the function of piece mobilization thus giving a holistic definition of the opening stage.

We should also note that the structural aspect of the game does not determine either the function or the process: it merely brings greater degree of correlation between them.

The endgame stage is characterized by a small number of pieces on the board. This obviously simplifies the task of predicting the "final position" since the number of possible variations is not so large.

A final position, i.e a position which leads to an immediate mate shall be termed terminal position.[94] The presence of terminal positions allows for the implementation of backtracking algorithms which start at the final position and work back to the initial ones. Here, the actual terminal position may have a direct impact on the criterion associated with each intermediate position. Since the number of terminal positions in chess is enormous, an algorithm is basically forced to move "forward" from the current to the final (relative to a given task) position. This gives rise to a very difficult problem of formulating a local criterion to evaluate intermediate positions.

These features of endgame stage reveal why this stage can be handled by an optimization algorithm[95] which ensures that a terminal position is reached in a minimum number of moves.[96]

Nevertheless, terminal positions can be set up even at the beginning of the game when there is a large number of immobile pieces. Example of this is the so called "Fool's mate": white to move and mate in three. In reality these considerations apply only to the beginners. At more advanced levels of play, terminal positions are so remote from the initial one that they defy clear-cut formulation not only in the opening but also in the middlegame.

---

[94] Terminal positions can occur at an earlier stage - a state which leads inevitably to a mating attack. This feature is incorporated into some computer chess programs which warn the opponent of a mate in a certain number of moves.

[95] "...while in the middlegame players' plans are largely determined by their tastes and fantasy, in the endgame they are determined bu the position in question. Everyone must more or less follow the same plan which is typical for a given position independently of his style or tastes. In fact, many endgame positions are nothing else than logical problems sometimes having a unique solution." (Averbasch, U., Things to Know About the Endgame, Moscow,1979).

[96] This may affect the "fifty move" rule which states that if no exchange of pieces takes place or no pawns are advanced then one side can claim a draw.

So, we have looked at the three stages of the game from the systems point of view. I recommend that the reader consult literature on this subject to fully appreciate the game (a partial list is given in references to Table 7). These books contain very interesting ideas on the various stages of the game. Unfortunately, these ideas are not presented systematically: the role of each aspect in the analysis is not stressed; its relation with other aspects of the game is not incorporated into the definitions of various stages. To illustrate this point, I want to quote from C. Shannon, the founder of computer chess. The quote contains a very concise description of the three stages of the game.

> "A game of chess can be divided into three phases, the opening, the middle game, and the end game. different principles of play apply in the different phases. In the opening, which generally lasts for about ten moves, development of the pieces to good positions is the main objective. During the middle game tactics and combinations are predominant. This phase lasts until most of the pieces are exchanged, leaving only kings, pawns and perhaps one or two pieces on each side. The end game is mainly concerned with pawn promotion. Exact timing and such possibilities as "Zuzwang", stalemate, etc. become important."[97]

Finally, concerning the level of development of different stages of the game.

> "Progress in present day general theory [of chess - A. K.] has been primarily in the opening stage. The ever growing amount of chess literature, "the information boom", has lead to the theory of opening stage "outrunning" other stages of the game. Study of openings is being conducted on all fronts, both in depth and in width. There appeared a breed of fanatics theoreticians who like moles dig deeper and deeper occasionally reaching beyond the middlegame and into the endgame.
>
> Evolution of opening theory has resulted in "real" game starting somewhere around move 20 or even 30. All previous moves are well known in advance. In these sometimes very sharp situations the decisive factor is frequently the player's book knowledge rather than his abilities.
>
> ...We should note that at the present time the middle game is the most underdeveloped area of chess theory. Numerous positions classified one way or another are of little help when it comes to real theory of the middle game, a general high quality theory which would really help orient practical players in the immense ocean of the middle game."[98]

---

[97] Shannon,G., "Programming a Computer for Playing Chess", The Philosophical Magazine, vol.XLI, 1950, p. 265.

[98] Averbasch, I., "Chess Today", in a book by Averbasch, I. and Talmanov, M., World Championship Match Karpov vs. Kasparov. Moscow: Fizcultura i Sport,1986.

Chapter 5. Styles of the game

§1. The subjective and the objective in chess: aesthetic and scientific methods.

Chess practise has shown that one can play only so well just by studying and playing the game. Further progress requires creative abilities. These abilities can be developed if a predisposition toward them is there. Creative abilities are crucial in chess. A youngster who lacks practical experience but possesses creative abilities can achieve very impressive results (the situation is similar in mathematics, sports, art but not in physics, social sciences, history, etc.). If a player does have creative abilities it would be interesting to see them at work, to see how they relate to the various methods of play, especially the combinational and the positional methods. Here, I want to discuss one aspect of this problem which in my opinion is of general methodological significance.

Combinational style of chess requires that a player have the talent to discover a combination, and not only its ultimate outcome but also a program of its realization. Therefore, having discovered a combination and having put it in more or less precise terms a player can verify it (sometimes with his assistants) before playing it in the actual game. Checking a combination for holes can produce absolutely unambiguous results, at least in the probabilistic sense of the word.

Positional play is a totally different story, especially at its non-goal directed stage or if it revolves around the accumulation of small advantages. First of all, a player can improve his position by introducing new positional parameters together with his individual scale of values based on his experience, knowledge, and abilities. A player possessing these abilities is in the same situation as a combinational player.[99]

Moreover, player's subjective experience in playing certain type of positions is not the only factor which determines his evaluation of a given position. His judgment also depends on his experience in realizing various positions, i.e. at transforming them into a combination. But what is the right time for this transformation? At what point of the game does a player, given his experience and abilities, start to look for potential combinations?[100] Here, the subjective factors prevail. One factor is player's confidence that given a certain position he will be able to squeeze most out of it. Under these circumstances, a positional player will be unable to demonstrate to his assistants a clear-cut way to realize his positional advantage: it is subject to his expertise, creative abilities, and originality.

---

[99] "In the end, all my valuations originate from my experiences: my first losses and wins which gave me pain and joy; my first draws that called forth in me a variety of sentiments; my first analysis, which was crude and faulty. From then on I valued and continue to value; and with practice I became capable of more exact valuation. And from this rough material is generated, by continued trial and intelligent criticism, the series of valuations by which the master arrives at his conclusions." "These valuations, and this is Steinitz's most important idea are supposed to guide us. They are like a compass for the sailor swimming in an ocean of combinations." (Lasker's Manual of Chess, Philadelphia: Mc Kay Co. 1947, p.191 and Lasker, E., Lasker's Chess Manual. Moscow: OGIZ, Fizkultura i Turizm, 1937, p.194).

Therefore, positional judgements can be controversial while a combination - its end result and its soundness can be judged by the experts unequivocally.

Subjectivity and objectivity in the styles of play correspond to scientific and aesthetic methods respectively. Which of them dominates depends on the stage of development. Let me clarify this statement.

It seems to me that the notion of the value of a position is nothing else than an aesthetic evaluation of its beauty. The word beauty does not apply to the end result of a combination since victory may be close at hand due to big material advantage. Actually, there are two factors determining the beauty of a position: on the one hand, a high value assigned to a given position merely signifies its greater predisposition towards future development. On the other hand, position's evaluation is not sufficient for the program to link constructively the material available on the board with the outcome of the game.[101]

So far, we have examined the combinational and the positional styles of play from the standpoint of the end result. Now, let us view them from a different perspective - the methods used in each stage to achieve the goal. What delights us in a combination is the unexpectedness of the moves. It is the great weight assigned to such parameter as "deep understanding" and the drastic departure from the orthodox and the stereotypical that induces in us this feeling of beauty of the combination. We can also be certain of the soundness of the combination initiated by the player's creative genius for it can be checked carefully and unambiguously. Positional style of play geared towards achieving a good position frequently does not impress us with its paradoxical moves. Positional moves, especially in non-goal directed problems or goal directed problems centered around the accumulation of small advantages, do not go against the grain of our common sense, especially that of professionals for whom positional style has become second nature, their traditional mode of thinking.

So, in positional chess beauty manifests itself primarily in the structure of the position. Its unorthodox nature represents one of the preconditions for future development. Beauty of a combinational style lies in the method of play which departs from stereotypical understanding.

Now let us ask ourselves the following question: can the category of beauty used in evaluating a position be applied to the combinational style of play? One of the most vulnerable aspects of a combinational style is formulating a goal. Goal formation which makes combinational style so attractive is at the same time the most obscure phenomenon: we frequently do not know what motivates players in their choice of a particular goal. It seems that target selection (setting a goal) is directly related to finding a "weak link" in the position of your opponent. Could it be that the weak link is less beautiful or even ugly? Beauty is not an absolute - it has gradations. A system as a whole can be beautiful and yet have some ugly parts. They detract from the beauty of the whole and their "development" can be very detrimental to the system as a whole. Therefore, the initial step in our search for a combination should be analysis of a position typical of positional style in order to pinpoint the least beautiful or perhaps even ugly spots in the opponent's position. For instance, one can start by ranking opponent's pieces based on their unconditional values supplemented with positional characteristics associated with each piece. Does not this total sum reflect the degree of beauty? And does not its negative value resulting from the prevalence of negative positional parameters indicate the extent of ugliness?

Of course, ranking opponent's material is only the first step. A player then faces a very non-trivial task of picking a specific target based on the current position on the board, i.e. he has to weight the benefits associated with achieving a particular goal against the expenditures for its realization. Is not this how chess players think on the intuitive level when they select a target? Do not animals behave in the same manner when hunting for prey?[102]

---

[100] M. Tal advances some interesting considerations in his account of Karpov-Rashkovskii game: "Karpov has played very consistently throughout the game but being strategically inclined towards king's Indian defense he had overlooked the opportunity to end the game thirty moves earlier.

...This experience came in handy. The third game of the Karpov-Spassky match player in half a year witnessed the same opening set up for black chosen by the former champion. Karpov won the game convincingly."

Tal, M., "All is ahead". Introduction to A. Karpov's book Selected Games 1969-1977. Moscow: Fizcultura i Sport,1978.

[101] I think that the concept of beauty in chess as seen by A. Roizman is quite similar to my own ideas on this subject. "Beauty in chess is really a multi-dimensional concept. Experts will speak with understanding of the beauty of say a positional maneuver or endgame technique. But nothing can really compete with the beauty of a combination. They are unexpected, effective, and seem very paradoxical. At the same time, they are very logical at some deeper level. All this makes a strong impression on the lovers of this ancient game. Another important factor is that combinations, because they are forced, are accessible to ordinary players." Roizman, A. J., Chess Miniatures, Minsk: Polymia, 1978, p. 3.

§ 2. A sacrifice in chess

The difference between the subjective and the objective in chess becomes very vivid when we look at a sacrifice, especially when we compare positional sacrifice, the most difficult one to judge, with a combinational sacrifice.

History of chess provides us with a unique opportunity to follow the evolution of the idea of a sacrifice, the interplay between its costs and benefits and its role in the most general case of creation of a position. It seems that chess players go through at least four stages in their understanding of the concept of a sacrifice. A beginner strives to gain at each step of the way, he never makes a sacrifice.[103]

---

[102] It seems to me that something similar takes place when the category of beauty is used in mathematics. In proving theorems, mathematicians often face the same sort of problems as do chess players: an exact algorithm is not available so we are forced to prove the hypothesis from the beginning and not from the end. The tree of "possible moves" grows very quickly preventing us from linking the current state with the end result. Therefore, we must be able to evaluate a position - an intermediate stage, a stepping stone for future development. It is reasonable to think many good mathematicians are guided by the criterion of beauty. But this is a rather trivial statement. What is less trivial is employing the notion of beauty to check if a proof is really sound. Looking at the relationship between human patterns of thought and computer idols, a professor from Columbia University B. G. Moishezon noticed this very interesting phenomenon. "...every professional mathematician whose work always entails finding mistakes knows that using a purely logical approach, i.e. checking each step one a time, does not lead to success. Mathematicians discover mistakes when they step aside or go back, i.e. when they use informal methods, when they compare the results to be checked against other information available to them. Only by comparing, deliberating, and going back and forth can a mathematician find the mistakes." Moishezon, B. G., "On Human Thinking and Computer Idols", Vremia I My, N°86, 1985, p. 82.

I agree with B. Moishezon's statement. I just want to make one method for uncovering mistakes in mathematical theorems more concrete.

At first glance, it seems that to prove the soundness of a theorem we just have to check each step for logic. Why then is it so difficult to pinpoint mistakes in the proofs? Why must decades sometimes pass before a logical mistake in the proof is discovered? What happens is that our mind is unable to check each little step for all details, especially its connection with parallel results achieved a few steps earlier. Thus, we are forced to approach a proof positionally. In proofs saturated with statements (material) and their interrelations (positional parameters) we must seek out those which are potentially liable to lead us in the wrong direction. Here, the primary candidates for a possible mistake are the ugly parts of the proof.

[103] This kind of immaturity in chess is typical of the behavior of some adults too. They want immediate results and are thus unwilling to make sacrifice today for a gain in the future.

[104] Spielman, R., Theory of Sacrifice, Moscow: Fizcultura i Turizm, 1936.

The next level of play involves the idea of a "pseudo" combinational sacrifice.[104] Here, a player is willing to make a sacrifice if the entire chain of moves of the combination which results in material gains (or an overall victory) is 100% certain.

At the next stage of the learning process the idea of a combinational sacrifice is supplemented by a probabilistic evaluation of the outcome. Naturally, carrying out such a combination requires more skill on the part of the player.

Finally, by the time a player has reached the fourth stage he is willing to make a sacrifice for the sake of improving his position. This kind of a sacrifice is called positional sacrifice.

Lasker, E., 1937 defines positional sacrifice in the following way:

> "Nowadays, a sacrifice based on the evaluation of position is called a "positional sacrifice". For instance, a pawn is sacrificed for the sake of development, a piece is given up for two pawns and an attack, and a queen is sacrificed for a rook and a bishop - all this for the sake of "position". These are not combinational sacrifices for the network of possible variations is too large, but they are surely the forerunners of combinations. And always is a positional sacrifice a sign of character and ability to think independently." (p.235).[105]

Positional sacrifice as well as positional style in general is a more recent development than a combinational style of play. Positional sacrifice does not preclude a combination, in fact they complement each other. Sometimes, when the possibilities of a combinational sacrifice have exhausted themselves it transforms into a positional sacrifice. This is what happened to king's gambit as played by white. Falkbeer's countergambit and other active modes of play for black have lessened the popularity of king's gambit.

> "It is presently played in the positional style, without ambitious attacks and sacrifices typical of the past".[106]

In making a positional sacrifice a player does not know exactly how the game will develop. The greater is our ability to foresee the consequences of a positional sacrifice

---

[104] Spielman, R., Theory of Sacrifice, Moscow: Fizcultura i Turizm, 1936.

[105] An Austrian grandmaster R. Spielman also gave a definition of a positional sacrifice: "A chance to convert power into material and material into power is a wonderful feature of the game, may be this is their greatest secret. "

Commenting on this quote from Spielman, P. Kondratiev writes that: "Subjecting myself to the dangers of analogies and metaphors, I should say a positional sacrifice exemplifies the law of the conservation of chess energy." Kondratiev, P. E., Positional Sacrifice, Moscow: Fizcultura i sport, 1983, pp. 3-4.

Much of our further discussion of positional sacrifice is based on Kondratiev's book, 1983.

[106] Roizman, A. J, Chess Miniatures, Minsk, Polymia, 1978, p.5.

(more precisely, to see the advantages of making a sacrifice) the closer it is to a combination.

So, a player is really uncertain of how and when a positional sacrifice will justify itself. In fact, a player making a sacrifice must be confident that his knowledge and talent is sufficient to realize his positional advantage at some future point of the game.

> "A player who decides to make a positional sacrifice bears additional responsibility: his play will have to be very forceful and precise. "Second rank" moves that are easier to find and that often turn out o'kay in normal positions will probably not do. They might lead to a loss of positional advantage and the initiative while the material advantage still belongs to the opponent."[107]

All this means that one must play at about master's level before he can afford to make a big material sacrifice for the sake of positional advantage. A combinational sacrifice determines the course of the game (completely or partially); an advantage gained from a positional sacrifice must be realized under the conditions of uncertainty. This points to the fact that a positional sacrifice is accessible only to players with certain innate abilities. Players differ not only in the amount of chess knowledge they had received or who their teachers were but also in the methods of play which, given our current level of knowledge and pedagogical skills in presenting the material, cannot be assimilated by everyone who wishes to use them (here, I do not mean to rule out the possibility that as our knowledge becomes deeper and teaching methods improve anybody can be taught to do anything).

In digression 3 I want to use an example of human patterns of behavior to illustrate the concept of a sacrifice in chess. For me the value of a particular point of view is revealed by whether or not it leads to conclusions *opposite in sign* to those arrived at on the basis of traditional methods used in that field, i.e. every day common sense.

Digression 3.

The notion of a positional or a combinational sacrifice frequently comes into play when a person is faced with having to make a decision .

I remember a conversation I had with a recent immigrant from the Soviet Union. A gifted engineer with the knowledge of English he received a decent job and good salary upon coming to the United States. After some years past his son was accepted to a private university. It was hard to pay for the son's education since much of the salary was spent on everyday things and their savings were negligible. This immigrant family had relatives in the U.S.: two older childless women who belonged to the lower upper class. (They were able to maintain their previous standard of living by living of interest on the interest gotten from their savings). These women maintained a rather close relationship with the immigrant family. Yet they offered no money to help finance the boy's education. So, this engineer complained to me about the *stupidity* of his relatives. He thought that by giving his family the financial help that they needed his relatives could have greatly improved their relations with his family and at the same time not suffer toc much financially.

I disagreed telling him that he really expects too much from his American relatives in the way of "understanding life". In other words, he demands that they make material sacrifice for the sake of improving positional parameters, i.e. improving relations with their immigrant relatives which may or may not help them in an indeterminate future. Here, help could have come in the form of warm feelings that cannot be bought at any price.

It seems that these American relatives of his decided that this kind of positional sacrifice is not for them. Money can buy them almost "anything" at their old age (medical care, a good retirement home, etc). Potential help and warmth from their immigrant relatives in the future was rather uncertain and far removed. Moreover, it takes a rather sophisticated individual to realize a positional sacrifice in personal relationships. Otherwise, an individual may suffer not only from being unable to realize a better position achieved as a result of a positional sacrifice but also, as we know from experience, from the deterioration of his relationship with those to whom he had extended his help. The American relatives were not sophisticated at all. Yet they were *smart* enough to understand their inability to realize the advantages of a positional sacrifice. They were like a clever intermediate level chess player who understands the inexpediency of committing to a positional sacrifice in his own games.

§ 3. *Logical and historical development of styles of play*.

Proceeding from the general to the particular, the logical sequence in the evolution of styles of play can be represented in the following way. At the start of the game, we set up a strategic goal-directed protractive positional problem of strategic variety which is followed by a tactical one belonging to the same class. Initially, this leads us to a compositional problem and then to a combinational one. We then move to a non-goal directed search problem. After that come the reactive methods of play.

The history of chess and the process of learning it follow the same above sequence except in reverse.

---

[107] Kondratiev, P. E., *Positional Sacrifice*, Moscow: Fizcultura i sport, 1983, p. 4.

Beginners, for the most part, play a combinational style of chess. They learn about different recommendations on how to play the game (see aggregate recommendations - heuristics). After that they face a very difficult question: "What do I do next?" They want to avoid inconsistencies inherent in these recommendations and seek ways to make their play more coherent. This leads them to think that material superiority is a precondition for victory. In fact, some players recommend that a direct attack against enemy king be initiated right in the middle of the game.[108]

Having mastered the combinational style of play a player can begin to assimilate the positional style.[109] This approach makes sense for without the ability to convert positional advantage into a combination it makes no sense to strive for positional superiority especially when it requires a material sacrifice.

There is a certain parallel between the process of learning and the development of the methods of play in general. Reactive methods based on personal wisdom prevailed in the past. This pertains not only to isolated rules but also to the entire games played by the great masters. Next came the combinational style which dominated the game until the end of the nineteenth century. The play aimed at gaining a large material advantage through beautiful combinations improvised by the players. Achieving this goal virtually ensured that the opponent's position would crumble very quickly. Usually the game did not last very long - about 20-25 moves. It seems that A. Andersen (1818-1879) represents the peak of combinational style.

Another great player P. Morphy (1837-1884) was first to introduce an innovative positional approach which violated the then prevalent style of play used even by the greatest masters of the day.

Let us note that before Morphy the games of chess masters contained an intermediate stage between opening and combination which aimed at loosening up the position. A sacrifice or some other sharp move produced tension on the board, and although the future course of the game remained unclear, a search for combination was thus made possible.

According to Jacob Neishadt[110],

> "The masters of combinational school, (this is partly true of Anderson in the 1850's) in opening up the game, often did not know how the play would continue. They

---

[108] Capablanca, H-R, Chess Manual, Moscow: Fizcultura i sport, 1975, p. 36.

[109] "One should master the combination style before attempting to play positionally." Reti, R., A Modern Manual on Chess, Moscow: Fizcultura i sport, 1981, p. 6.

[110] Neishadt, J. Uncrowned Champions. Moscow: Fizkultura i Sport, 1975. p. 286

> were convinced that in a great majority of complicated positions - those in which their pieces were not obstructed by "damned" pawns - they would find a profitable combination. Morphy was the one who understood the error of this approach, and would open up lines only when he had an advantage in development."

In particular, Morphy went after small positional advantages; among other things, he was willing to trade heavy pieces so important in carrying out combinations for the sake of gaining a small positional advantage. Morphy was very successful for he was able to conver his positional advantage into a combination. According to a great chess player R. Ret Morphy became the most famous "of all the people who ever played the game".[111]

V. Steinitz (1836-1900) was able to conceptualize positional style of play thus making it accessible to other players. E. Lasker writes

> "How novel, how surprising, how opposed to every sentiment of his time the conceptions of Steinitz must have been...
>
> ... The main principle in Stenitz's conception could be expressed in the following way: a master's plan should always be based on his evaluation of the position.
>
> The style with which Steinitz began his career was strange and narrow, but it was the style predominating in his time... The play was wholly dominated by the feverish desire to make a rush against the hostile king and to this end furiously to assail the obstructions, regardless of the sacrifices required.
>
> Surely, Steinitz's heart beat when for the first time the thought came to him that the master should not look for winning combinations, unless he believed, unless he could prove to himself that he held an advantage. That meant making no attempt at winning in the beginning of the game. And since Steinitz lived in a milieu where to play to win right from the start was considered the only honorable course to take, this thought must at first have had a timid reception in his mind and a hard time establishing itself."[112]

Besides developing the techniques of positional play E. Lasker introduce philosophy and aesthetics into the game. Subsequently positional style was perfected by R Capablanca, A. Nimzovitch, and other chess greats. At the present time, positional style is the leading and quite diverse style of play. It includes non-goal as well as different types of goal directed positional problems .

As E. Lasker point out opposing different methods of play is characteristics only o beginners. In masters' play the two styles complement each other. The linear picture of the game's history presented above is really more complicated. At different stages of the game one style of play is combined with another which by itself may belong to an earlier period of the game's evolution. This coalition often represents a very powerful combination.

---

[111] Ibid., p.16.

[112] Lasker, E., Lasker's Chess Manual, Moscow: OGIZ, Fizkultura i Turizm, 1937, pp. 193-195 and Lasker, E., Lasker's Manual of Chess, Philadelphia: Mc Kay Co., 1947, p.197.

Things will become more clear if we view the process and the structural approaches to the classification of various stages together.

Middle game is the most important example in this connection. It is at this stage that we face some of the most perplexing problems whose solution has a general methodological significance.

Designing procedures for formulating local problems presents the most difficult task. Indeed, a large number of rather mobile pieces and an astronomical number of possible variations in the middle game raises two issues: on the one hand, unlike the end game, there is no way to reach the final goal, i.e. to formulate a terminal position and an optimization procedure to reach it. On the other hand, the play cannot be reduced to a reactive method based on prior experience (as is the case in the opening).

Looking at the chess game from a global perspective we see that each stage of the game embodies different methods of play including the reactive method. It is interesting to note that the reactive approach characteristic of the early history of the game may come into play at the game's more mature phase. But here these techniques would be much more sophisticated being applied in combination with search procedures.

Church, R. and Church, K. note in this respect:

> "The most reliable method for choosing a move is based upon complete pattern recognition. Particularly in the openings, a chess player has associated many positions with specific moves. When the player recognizes the position as one which has been previously encountered, the associated move can be executed without any analysis. Such moves can be made quickly and accurately.
>
> In the middlegame, it is not likely that the player will have encountered the complete position previously, but aspects of the position may include familiar patterns. For example, certain tactical elements (e.g., pin, discovery, x-ray, skewer, overworked piece, trap, and fork), certain structural features of the pawn formation, and other features of the position may be immediately recognized as patterns. These patterns will not be associated with specific moves, but with goals that the player might attempt to realize, and various plans to achieve these goals.
>
> The recognition of winning and drawing endgame patterns may be an essential factor in the choice of moves at any stage of the game. A player who attempts to work out a complicated rook and pawn endgame that his opponent recognized would be at a substantial disadvantage. As in the case of middlegame features, endgame patterns will normally be associated with goals and plans rather than specific moves. The treatment of the legal move as the elementary unit of chess thinking is concrete but undoubtedly misguided."[113]

M. Botvinnik stressed the idea of using different methods of play at different structural stages of the game. Here is what he writes:

> "When a chess master plays a game he uses historical experience (his own and the experience of others) in four different ways:
>
> 1. In parrot fashion. This is characteristic of the opening. The theory of the opening is not subject to judgment in any of its parts, and the master makes his moves in the opening without inserting himself into the process, i.e., he moves as a parrot talks.
>
> 2. By the handbook method. In playing a game, the master seeks out, in his store of accumulated knowledge, the exact position occurring in the search. This library position has a score; since the positions coincide, the score of the variation is known, and the variation itself may be cut off. This usage of historical experience is characteristic of the endgame, but may be successful in the opening as well, when moves are transposed.
>
> 3. By the outreach method. This is based on an attempt to reach positions that the handbook method can use. The master looks up library positions having a favorable outcome and resembling[114] the search position. (See the Glossary of Terms for a definition of nearness.) Having found such positions, the master constructs his search so as to get them into the search tree if possible. Then he relies on method 2. Therefore this method is also characteristic of the endgame and, perhaps, of the opening.
>
> 4. By the associative method. This is based on a partial congruence of the position in the search with positions in the library. The master seeks out a fragment of a library position, i.e., a small group of pieces whose action has led to success. If the search position contains the same group of pieces in the same constellation, he includes the group in his search, in order to see whether it will lead to success in the current case. If the fragment has often been successful in the past, it is apt to succeed again. This method determines a direction for the search and on the average it saves resources while the search is under construction. It would appear that the associative method presents the only way to apply historical experience to the middle game and to complex endgames."[115]

---

[113] Church, R. & Church, K., "Plans, Goals, and Search Strategies for the Selection of a Move in Chess", Chess Skill in Man and Machine, ed. by P. Frey. New York: Springer-Verlag, 1977, pp. 136-137.

[114] Maximum allowable (D_{max}): the largest number of discrepancies between two positions equal in material that is permitted if the two positions are to be regarded as near each other;
between two positions equal in material (DAB) : the sum of the lengths of the trajectories leading from one (initial) position to the other (library) position." (p.146).

[115] Botvinnik, M., The Solution of Inexact Problems. Moscow: Soviet Radio, 1979.

I leave it up to a professional chess player F. Reinfeld to sum up our discussion of the styles of play in chess. While I do not agree with him completely he managed to express these ideas better than I.

> "The style is the man himself," says Buffon, and nowhere is the famous phrase (Le style est l'homme meme) more appropriate than in chess. Many men, many styles; and what is chess style but the intangible expression of the will to win? The universe of the chess master is not without its grimmer aspects, for it is a world of dog-eat-dog. Beauty in chess (like virtue) is its own reward; it is only the incidental by-product of relentless struggle.
>
> In such an atmosphere, the quality of objective appreciation is not seen too frequently. Yet differences in style may produce queer paradoxes".[116]

---

[116] [illegible] Nimzovich the Hypermodern Philadelphia, David McKay Co., 1948, p.23.

Conclusion

The game of chess models some very important phenomena taking place in the world around us. Chess is a model of competition based on equal starting conditions of all the participants with the outcome depending exclusively on their respective abilities to solve non-trivial problems. At the same time, many rules in chess are rooted in other moral statutes reflecting the social structure of society.

Despite its simplicity, the game yields to an experimental verification of a number of very important procedures aimed at achieving a certain result. The need to develop these procedures stems from the impossibility of linking the initial state with the end result by carrying out an exhaustive search of all possible variations and then designing an optimization algorithm to determine the best course of action.

One of the greatest achievements of chess theoreticians has been the development and synthesis of various techniques and methods to help guide our local behavior. Analysis of local problems has revealed a great variety among these methods. They can be viewed from different perspectives and within a hierarchical scheme. This enables us to combine different approaches to help solve local and global problems: problems limited to one move as well as problems encompassing an entire stage of the game (such as an opening, a middlegame or an endgame). As human experience has shown it is the synthesis of this manifold of local problems rather than its unification that is a crucial prerequisite for designing an effective algorithm of the game.

Of all the various methods of play, the two that are most noteworthy are the reactive and the search methods. I think that the greatest achievement of the game has been the discovery of search procedures aimed at creating good positions - the forerunners of combinations. No other field was able to develop the concept of integrating parameters in such a complete and precise fashion. Parameters which enable us to link one position with another in the absence of a procedure used to estimate completely and consistently the impact of a given position on the future flow of the game. Parameters of integration include both primary and conjugate parameters: material and its evaluation, positional parameters and their respective weights. Integrating parameters represent arguments of the weight function used to estimate the value of a position. A very special feature of this function is that it incorporates material as well as positional parameters as independent variables.

Numerous ways of modifying the structure of the weight function bring diversity in the kinds of search problems we have to deal with. Goal-directed problems are a very important category. These problems include a specific goal of reaching a certain position conceived beforehand. This class includes an important subset of combination-like objectives aimed at gaining material superiority with the program of achieving this goal known.

Despite all the techniques that have been developed by the players and the extent to which human experience has been formalized, subjective understanding still plays a tremendous role in the game. It manifests itself most vividly in positional style when a player has to judge the beauty of a position based on his experience and understanding. It is the uncertainty about the future progress of the game that makes positional judgments so subjective: indeed, the position in question is going to be played by a particular player and no one can tell ahead of time what the outcome might be. A sacrifice in chess puts the objective and the subjective aspects of the game in relief. It is also quite feasible that uncovering the least beautiful or even ugly spots in opponent's position is very important in goal formulation, especially that of capturing the material.

In spite of everything that has been accomplished in chess including the development of computer chess, the game still maintains its original vitality. It also yields to interesting experiments in solving non-trivial problems called by M. Botvinnik "inexact problems." Chess players have accumulated a lot of experience in the game. We were able, at least partially, to convert this experience and various techniques into formal procedures used in computer algorithms. Still, much of what has been accomplished in chess is still very far from being translated into a formal language. This applies equally well to the methods of solving similar types of problems arising in other fields. Some major puzzles in chess are still unresolved: formalization of the goal set up by within a local problem, the set of positional parameters and their weights, aggregation of different parameters, etc.

Particularly troublesome is the problem of formalizing the learning process in chess: this is still the prerogative of man. The problem consists not only in designing formal algorithms to implement the already existing methods of defining local problems but in teaching the machine to perfect these methods and to invent new ones. In other words, we are talking about creating an algorithm of the game capable of improving. I use the term "improving" intentionally. It encompasses both external and internal sources of learning and promotes the idea of a machine being able to create new methods on its own accord.

Until now, the emphasis in algorithmizing the game of chess has been on designing chess programs from the outside by a team well-qualified individuals (players, mathematicians, programmers). The task of designing a program to improve the game algorithm from the inside was often reduced to the creation of self-learning algorithms, i.e. a second-level algorithms aimed at perfecting the first-level algorithm based on the information accumulated from playing the game.

This information is received from many sources: games played by human beings, by humans and machines, and games played by a machine with other machines or with itself.

A number of difficulties in designing a self-learning algorithm came to the surface. The major ones were: 1) setting up a standard positional problem, i.e. elaborating new positional parameters and improving upon their weight assignment, 2) lack of sufficiently sophisticated methods for setting goal-oriented local problems, 3) "overabundance" and the contradictory nature of the heuristics incorporated into the reactive methods of play. Judging by the amount of literature on self-learning algorithms[117] the progress in this area has been rather slow.

We should note another problem which lies outside the scope of the game itself. This is the creation of algorithms whose purpose is to develop new algorithms of the game rather than just improve upon the old ones.

It seems that one of the major obstacles in creating such algorithms or in improving upon the structure of the already existing ones (rather than just modifying the values of constant coefficients) is the fact that the learning process is not limited exclusively to chess.

Experience gotten from the game itself may be insufficient: what is important is the creative potential of the creator of the algorithms, both self-learning ones and those aimed at improving other algorithms.

> I want to quote E. Lasker's thoughts on this subject: "Steinitz did not deduce his concept from just studying master games; finding the basis for it in human practice in general, he later tried to reconcile it with a purely chess experience. His venture was crowned with success."[118]

There is no reason to doubt that eventually the game of chess will be formalized to such an extent as to make an algorithm which finds an optimal strategy quite feasible. Even prior to developing such an optimization algorithm we could come up with a program which would be stronger than the world champion. Progress in computer chess could contribute enormously to the development of human thought in general for the range of problems similar to chess is very wide. And just as people ride horses in the age of airplanes and automobiles they will continue to enjoy the game of chess in the age of computers.

---

[117] For example, Nitshe, T., "A Learning Chess Program", Advances in Computer Chess, ed. by M. R. B. Clarke, no.3, New York: Pergamon Press, 1982, pp. 113-120.

[118] Lasker, E., Lasker's Chess Manual. Moscow: OGIZ, Fizcultura i Turizm [illegible] p.248.

2-3

2

WHARTON REPROGRAPHICS

11

Programming a Computer for Playing Chess *.

By Claude E. Shannon,
Bell Telephone Laboratories, Inc., Murray Hill, N.J. †

[Received November 8, 1949.]

1. Introduction.

This paper is concerned with the problem of constructing a computing routine or "program" for a modern general purpose computer which will enable it to play chess. Although perhaps of no practical importance, the question is of theoretical interest, and it is hoped that a satisfactory solution of this problem will act as a wedge in attacking other problems of a similar nature and of greater significance. Some possibilities in this direction are :—

(1) Machines for designing filters, equalizers, etc.

(2) Machines for designing relay and switching circuits.

(3) Machines which will handle routing of telephone calls based on the individual circumstances rather than by fixed patterns.

(4) Machines for performing symbolic (non-numerical) mathematical operations.

(5) Machines capable of translating from one language to another.

(6) Machines for making strategic decisions in simplified military operations.

(7) Machines capable of orchestrating a melody.

(8) Machines capable of logical deduction.

It is believed that all of these and many other devices of a similar nature are possible developments in the immediate future. The techniques developed for modern electronic and relay type computers make them not only theoretical possibilities, but in several cases worthy of serious consideration from the economic point of view.

Machines of this general type are an extension over the ordinary use of numerical computers in several ways. First, the entities dealt with are not primarily numbers, but rather chess positions, circuits, mathematical expressions, words, etc. Second, the proper procedure involves general principles, something of the nature of judgment, and considerable trial and error, rather than a strict, unalterable computing process. Finally, the solutions of these problems are not merely right or wrong but have a continuous range of "quality" from the best down to the worst. We might be satisfied with a machine that designed good filters even though they were not always the best possible.

---

* First presented at the National IRE Convention, March 9, 1949, New York, U.S.A.

† Communicated by the Author.

The chess machine is an ideal one to start with, since: (1) the problem is sharply defined both in allowed operations (the moves) and in the ultimate goal (checkmate); (2) it is neither so simple as to be trivial nor too difficult for satisfactory solution; (3) chess is generally considered to require "thinking" for skilful play; a solution of this problem will force us either to admit the possibility of mechanized thinking or to further restrict our concept of "thinking"; (4) the discrete structure of chess fits well into the digital nature of modern computers.

There is already a considerable literature on the subject of chess-playing machines. During the late 18th and early 19th centuries, the Maelzel Chess Automaton, a device invented by von Kempelen, was exhibited widely as a chess-playing machine. A number of papers appeared at the time, including an analytical essay by Edgar Allan Poe (entitled Maelzel's Chess Player) purporting to explain its operation. Most of these writers concluded, quite correctly, that the Automaton was operated by a concealed human chess-master; the arguments leading to this conclusion, however, were frequently fallacious. Poe assumes, for example, that it is as easy to design a machine which will invariably win as one which wins occasionally, and argues that since the Automaton was not invincible it was therefore operated by a human, a clear *non sequitur*. For a complete account of the history and method of operation of the Automaton, the reader is referred to a series of articles by Harkness and Battell in *Chess Review*, 1947.

A more honest attempt to design a chess-playing machine was made in 1914 by Torrès y Quévedo, who constructed a device which played an end game of king and rook against king (Vigneron, 1914). The machine played the side with king and rook and would force checkmate in a few moves however its human opponent played. Since an explicit set of rules can be given for making satisfactory moves in such an end game, the problem is relatively simple, but the idea was quite advanced for that period.

The thesis we will develop is that modern general purpose computers can be used to play a tolerably good game of chess by the use of a suitable computing routine or "program". While the approach given here is believed fundamentally sound, it will be evident that much further experimental and theoretical work remains to be done.

2. GENERAL CONSIDERATIONS.

A chess "position" may be defined to include the following data:—

(1) A statement of the positions of all pieces on the board.

(2) A statement of which side, White or Black, has the move.

(3) A statement as to whether the kings and rooks have moved. This is important since by moving a rook, for example, the right to castle on that side is forfeited.

(4) A statement of, say, the last move. This will determine whether a possible *en passant* capture is legal, since this privilege is forfeited after one move.

(5) A statement of the number of moves made since the last pawn move or capture. This is important because of the 50 move drawing rule. For simplicity, we will ignore the rule of draw after three repetitions of a position.

In chess there is no chance element apart from the original choice of which player has the first move. This is in contrast with card games, backgammon, etc. Furthermore, in chess each of the two opponents has " perfect information " at each move as to all previous moves (in contrast with Kriegspiel, for example). These two facts imply (von Neumann and Morgenstern, 1944) that any given position of the chess pieces must be either :—

(1) A won position for White. That is, White can force a win, however Black defends.

(2) A draw position. White can force at least a draw, however Black plays, and likewise Black can force at least a draw, however White plays. If both sides play correctly the game will end in a draw.

(3) A won position for Black. Black can force a win, however White plays.

This is, for practical purposes, of the nature of an existence theorem. No practical method is known for determining to which of the three categories a general position belongs. If there were chess would lose most of its interest as a game. One could determine whether the initial position is won, drawn, or lost for White and the outcome of a game between opponents knowing the method would be fully determined at the choice of the first move. Supposing the initial position a draw (as suggested by empirical evidence from master games*) every game would end in a draw.

It is interesting that a slight change in the rules of chess gives a game for which it is provable that White has at least a draw in the initial position. Suppose the rules the same as those of chess except that a player is not forced to move a piece at his turn to play, but may, if he chooses, " pass ". Then we can prove as a theorem that White can at least draw by proper play. For in the initial position either he has a winning move or not. If so, let him make this move. If not, let him pass. Black is now faced with essentially the same position that White had before, because of the mirror symmetry of the initial position †. Since White had no winning move before, Black has none now. Hence, Black at best can draw. Therefore, in either case White can at least draw.

---

* The world championship match between Capablanca and Alekhine ended with the score Alekhine 6, Capablanca 3, drawn 25.

† The fact that the number of moves remaining before a draw is called by the 50-move rule has decreased does not affect this argument.

In some games there is a simple *evaluation function* $f(P)$ which can be applied to a position P and whose value determines to which category (won, lost, etc.) the position P belongs. In the game of Nim (Hardy and Wright, 1938), for example, this can be determined by writing the number of matches in each pile in binary notation. These numbers are arranged in a column (as though to add them). If the number of ones in each column is even, the position is lost for the player about to move, otherwise won.

If such an evaluation function $f(P)$ can be found for a game it is easy to design a machine capable of perfect play. It would never lose or draw a won position and never lose a drawn position and if the opponent ever made a mistake the machine would capitalize on it. This could be done as follows: Suppose

$$f(P)=1 \text{ for a won position,}$$
$$f(P)=0 \text{ for a drawn position,}$$
$$f(P)=-1 \text{ for a lost position.}$$

At the machine's turn to move it calculates $f(P)$ for the various positions obtained from the present position by each possible move that can be made. It chooses that move (or one of the set) giving the maximum value to f. In the case of Nim where such a function $f(P)$ is known, a machine has actually been constructed which plays a perfect game [*].

With chess it is possible, *in principle*, to play a perfect game or construct a machine to do so as follows: One considers in a given position all possible moves, then all moves for the opponent, etc., to the end of the game (in each variation). The end must occur, by the rules of the game, after a finite number of moves[†] (remembering the 50 move drawing rule). Each of these variations ends in win, loss or draw. By working backward from the end one can determine whether there is a forced win, the position is a draw or is lost. It is easy to show, however, that even with the high computing speeds available in electronic calculators this computation is impractical. In typical chess positions there will be of the order of 30 legal moves. The number holds fairly constant until the game is nearly finished as shown in Fig. 1. This graph was constructed from data given by De Groot, who averaged the number of legal moves in a large number of master games (De Groot, 1946, *a*). Thus a move for White and then one for Black gives about 10^3 possibilities. A typical game lasts about 40 moves to resignation of one party. This is conservative for our calculation since the machine should calculate out to checkmate, not resignation. However, even at this figure there will be

---

[*] Condon, Tawney and Derr, U.S. Patent 2,215,544. The "Nimotron" based on this patent was built and exhibited by Westinghouse at the 1938 New York World's Fair. impossible (5,300)

[†] The longest possible chess game is 6350 moves, allowing 50 moves between each pawn move or capture. The longest tournament game on record between masters lasted 168 moves, and the shortest four moves. (Chernev, *Curious Chess Facts*, The Black Knight Press, 1937.)

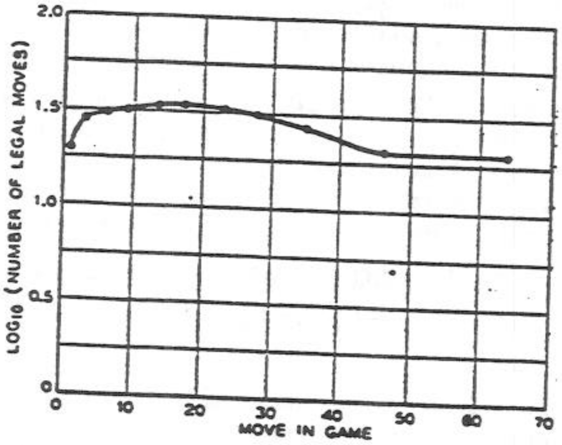
10^{120} variations to be calculated from the initial position. A machine operating at the rate of one variation per micro-microsecond would require over 10^{90} years to calculate its first move!

Another (equally impractical) method is to have a "dictionary" of all possible positions of the chess pieces. For each possible position there is an entry giving the correct move (either calculated by the above process or supplied by a chess master). At the machine's turn to move it merely looks up the position and makes the indicated move. The number of possible positions, of the general order of $64!\,|\,32!\,8!^2\,2!^6$, or roughly 10^{43}, naturally makes such a design unfeasible.

It is clear then that the problem is not that of designing a machine to play perfect chess (which is quite impractical) nor one which merely plays legal chess (which is trivial). We would like it to play a skilful game, perhaps comparable to that of a good human player.

A *strategy* for chess may be described as a process for choosing a move in any given position. If the process always chooses the same move in the same position the strategy is known in the theory of games as a "pure" strategy. If the process involves statistical elements and does not always result in the same choice it is a "mixed" strategy. The following are simple examples of strategies:—

Fig. 1.



(1) Number the possible legal moves in the position P, according to some standard procedure. Choose the first on the list. This is a pure strategy.

(2) Number the legal moves and choose one at random from the list. This is a mixed strategy.

Both of these, of course, are extremely poor strategies, making no attempt to select good moves. Our problem is to develop a tolerably good strategy for selecting the move to be made.

3. APPROXIMATE EVALUATING FUNCTIONS.

Although in chess there is no known simple and exact evaluating function $f(P)$, and probably never will be because of the arbitrary and complicated nature of the rules of the game, it is still possible to perform an approximate evaluation of a position. Any good chess player must, in fact, be able to perform such a position evaluation. Evaluations are based on the general structure of the position, the number and kind of Black and White pieces, pawn formation, mobility, etc. These evaluations are not perfect, but the stronger the player the better his evaluations. Most of the maxims and principles of correct play are really assertions about evaluating positions, for example :—

(1) The relative values of queen, rook, bishop, knight and pawn are about 9, 5, 3, 3, 1, respectively. Thus other things being equal (!) if we add the numbers of pieces for the two sides with these coefficients, the side with the largest total has the better position.

(2) Rooks should be placed on open files. This is part of a more general principle that the side with the greater mobility, other things equal, has the better game.

(3) Backward, isolated and doubled pawns are weak.

(4) An exposed king is a weakness (until the end game).

These and similar principles are only generalizations from empirical evidence of numerous games, and only have a kind of statistical validity. Probably any chess principle can be contradicted by particular counter examples. However, from these principles one can construct a crude evaluation function. The following is an example :—

$$f(P)=200(K-K')+9(Q-Q')+5(R-R')+3(B-B'+N-N')+(P-P')$$
$$-{\cdot}5(D-D'+S-S'+I-I')+{\cdot}1(M-M')+\ \ldots$$

in which :—

K, Q, R, B, N, P are the number of White kings, queens, rooks, bishops, knights and pawns on the board.

D, S, I are doubled, backward and isolated White pawns.

M=White mobility (measured, say, as the number of legal moves available to White).

Primed letters are the similar quantities for Black.

The coefficients ·5 and ·1 are merely the writer's rough estimate. Furthermore, there are many other terms that should be included*. The formula is given only for illustrative purposes. Checkmate has been artificially included here by giving the king the large value 200 (anything greater than the maximum of all other terms would do).

It may be noted that this approximate evaluation $f(P)$ has a more or less continuous range of possible values, while with an exact evaluation there are only three possible values. This is as it should be. In practical play a position may be an "easy win" if a player is, for example, a queen ahead, or a very difficult win with only a pawn advantage. In the former case there are many ways to win while in the latter exact play is required, and a single mistake often destroys the advantage. The unlimited intellects assumed in the theory of games, on the other hand, never make a mistake and the smallest winning advantage is as good as mate in one. A game between two such mental giants, Mr. A and Mr. B, would proceed as follows. They sit down at the chessboard, draw for colours, and then survey the pieces for a moment. Then either

---

* See Appendix I.

(1) Mr. A says, "I resign" or
(2) Mr. B says, "I resign" or
(3) Mr. A says, "I offer a draw," and Mr. B replies, "I accept."

4. Strategy Based on an Evaluation Function.

A very important point about the simple type of evaluation function given above (and general principles of chess) is that they can only be applied in relatively quiescent positions. For example, in an exchange of queens White plays, say, Q × Q (×=captures) and Black will reply P × Q. It would be absurd to calculate the function $f(P)$ after Q × Q while White is, for a moment, a queen ahead, since Black will immediately recover it. More generally it is meaningless to calculate an evaluation of the general type given above during the course of a combination or a series of exchanges.

More terms could be added to $f(P)$ to account for exchanges in progress, but it appears that combinations, and forced variations in general, are better accounted for by examination of specific variations. This is, in fact, the way chess players calculate. A certain number of variations are investigated move by move until a more or less quiescent position is reached and at this point something of the nature of an evaluation is applied to the resulting position. The player chooses the variation leading to the highest evaluation for him when the opponent is assumed to be playing to reduce this evaluation.

The process can be described mathematically. We omit at first the fact that $f(P)$ should only be applied in quiescent positions. A strategy of play based on $f(P)$ and operating one move deep is the following. Let $M_1, M_2, M_3, \ldots, M_s$ be the moves that can be made in position P and let M_1P, M_2P, etc. denote symbolically the resulting positions when M_1, M_2, etc. are applied to P. Then one chooses the M_m which maximizes $f(M_mP)$.

A deeper strategy would consider the opponent's replies. Let $M_{i1}, M_{i2}, \ldots, M_{is}$ be the possible answers by Black, if White chooses move M_i. Black should play to *minimize* $f(P)$. Furthermore, his choice occurs *after* White's move. Thus, if White plays M_i Black may be assumed to play the M_{ij} such that

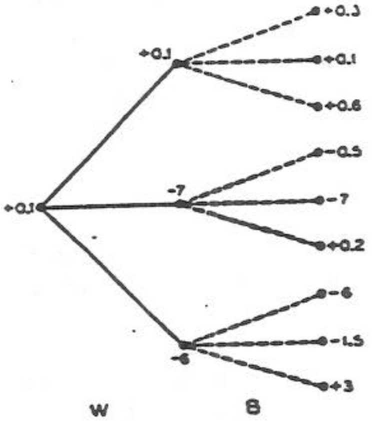
$$f(M_{ij}M_iP)$$

is a *minimum.* White should play his first move such that f is a maximum after Black chooses his best reply. Therefore, White should play to maximize on M_i the quantity

$$\min_{M_{ij}} f(M_{ij} M_i P).$$

The mathematical process involved is shown for a simple case in Fig. 2. The point at the left represents the position being considered. It is assumed that there are three possible moves for White, indicated by the three solid lines, and if any of these is made there are three possible moves for Black, indicated by the dashed lines. The possible positions after a White and Black move are then the nine points on the right, and the numbers are the evaluations for these positions. Minimizing on the upper three gives $+\cdot 1$ which is the resulting value if White chooses the upper variation and Black replies with his best move. Similarly, the second and third moves lead to values of -7 and -6. Maximizing on White's move, we obtain $+\cdot 1$ with the upper move as White's best choice.

Fig. 2.



In a similar way a two-move strategy (based on considering all variations out to 2 moves) is given by

$$\underset{M_i}{\text{Max}}\ \underset{M_{ij}}{\text{Min}}\ \underset{M_{ijk}}{\text{Max}}\ \underset{M_{ijkl}}{\text{Min}}\ f(M_{ijkl}\ M_{ijk}\ M_{ij}\ M_i\ P) \qquad . \ . \ . \ . \ (1)$$

The order of maximizing and minimizing this function is important. It derives from the fact that the choices of moves occur in a definite order.

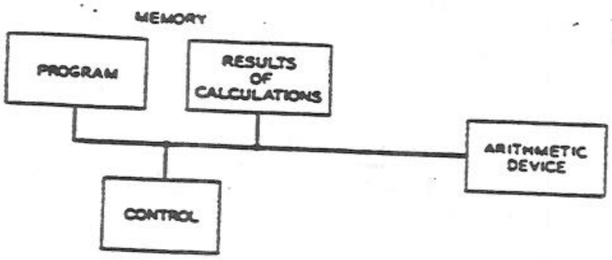
A machine operating on this strategy at the two-move level would first calculate all variations out to two moves (for each side) and the resulting positions. The evaluations $f(P)$ are calculated for each of these positions. Fixing all but the last Black move, this last is varied and the move chosen which minimizes f. This is Black's assumed last move in the variation in question. Another move for White's second move is chosen and the process repeated for Black's second move. This is done for each second White move and the one chosen giving the largest final f (after Black's best assumed reply in each case). In this way White's second move in each variation is determined. Continuing in this way the machine works back to the present position and the best first White move. This move is then played. This process generalizes in the obvious way for any number of moves.

A strategy of this sort, in which all variations are considered out to a definite number of moves and the move then determined from a formula such as (1) will be called a type A strategy. The type A strategy has certain basic weaknesses, which we will discuss later, but is conceptually simple, and we will first show how a computer can be programmed for such a strategy.

5. PROGRAMMING A GENERAL PURPOSE COMPUTER FOR A TYPE A STRATEGY.

We assume a large-scale digital computer, indicated schematically in Fig. 3, with the following properties :—

Fig. 3.



(1) There is a large internal memory for storing numbers. The memory is divided into a number of boxes each capable of holding, say, a ten-digit number. Each box is assigned a " box number ".

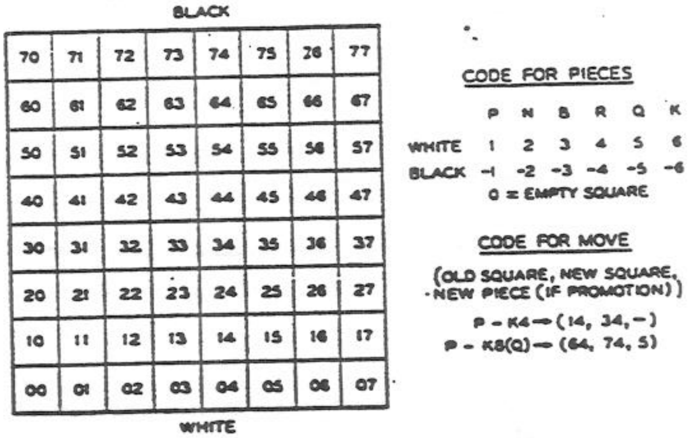
(2) There is an arithmetic organ which can perform the elementary operations of addition, multiplication, etc.

(3) The computer operates under the control of a " program ". The program consists of a sequence of elementary " orders ". A typical order is A 372, 451, 133. This means, extract the contents of box 372 and of box 451, add these numbers, and put the sum in box 133. Another type of order involves a decision, for example, C 291, 118, 345. This tells the machine to compare the contents of box 291 and 118. If the first is larger the machine goes on to the next order in the program. If not, it takes its next order from box 345. This type of order enables the machine to choose from alternative procedures, depending on the results of previous calculations. It is assumed that orders are available for transferring numbers, the arithmetic operations, and decisions.

Our problem is to represent chess as numbers and operations on numbers, and to reduce the strategy decided upon to a sequence of computer orders. We will not carry this out in detail but only outline the programs. As a colleague puts it, the final program for a computer must be written in words of one microsyllable.

The rather Procrustean tactics of forcing chess into an arithmetic computer are dictated by economic considerations. Ideally, we would like to design a special computer for chess containing, in place of the arithmetic organ, a "chess organ" specifically designed to perform the simple chess calculations. Although a large improvement in speed of operation would undoubtedly result, the initial cost of computers seems to prohibit such a possibility. It is planned, however, to experiment with a simple strategy on one of the numerical computers now being constructed.

Fig. 4.



BLACK

70	71	72	73	74	75	26	77
60	61	62	63	64	65	66	67
50	51	52	53	54	55	56	57
40	41	42	43	44	45	46	47
30	31	32	33	34	35	36	37
20	21	22	23	24	25	26	27
10	11	12	13	14	15	16	17
00	01	02	03	04	05	06	07

WHITE

CODE FOR PIECES

	P	N	B	R	Q	K
WHITE	1	2	3	4	5	6
BLACK	-1	-2	-3	-4	-5	-6

0 = EMPTY SQUARE

CODE FOR MOVE

(OLD SQUARE, NEW SQUARE, NEW PIECE (IF PROMOTION))

P - K4 → (14, 34, -)
P - K8(Q) → (64, 74, 5)

A game of chess can be divided into three phases, the opening, the middle game, and the end game. Different principles of play apply in the different phases. In the opening, which generally lasts for about ten moves, development of the pieces to good positions is the main objective. During the middle game tactics and combinations are predominant. This phase lasts until most of the pieces are exchanged, leaving only kings, pawns and perhaps one or two pieces on each side. The end game is mainly concerned with pawn promotion. Exact timing and such possibilities as "Zugzwang", stalemate, etc. become important.

Due to the difference in strategic aims, different programs should be used for the different phases of a game. We will be chiefly concerned with the middle game and will not consider the end game at all. There seems no reason, however, why an end game strategy cannot be designed and programmed equally well.

A square on a chessboard can be occupied in 13 different ways: either it is empty (0) or occupied by one of the six possible kinds of White pieces (P=1, N=2, B=3, R=4, Q=5, K=6) or one of the six possible Black pieces (P=−1, N=−2, . . ., K=−6). Thus, the state of a square is specified by giving an integer from -6 to $+6$. The 64 squares can be numbered according to a co-ordinate system as shown in Fig. 4. The position of all pieces is then given by a sequence of 64 numbers each lying between -6 and $+6$. A total of 256 bits (binary digits) is sufficient memory in this representation. Although not the most efficient encoding, it is a convenient one for calculation. One further number λ will be $+1$ or -1 according as it is White's or Black's move. A few more should be added for data relating to castling privileges (whether the White or Black kings and rooks have moved), and *en passant* captures (*e.g.*, a statement of the last move). We will neglect these, however. In this notation the starting chess position is given by:—

4, 2, 3, 5, 6, 3, 2, 4; 1, 1, 1, 1, 1, 1, 1, 1;
0, 0, 0, 0, 0, 0, 0, 0; 0, 0, 0, 0, 0, 0, 0, 0;
0, 0, 0, 0, 0, 0, 0, 0; 0, 0, 0, 0, 0, 0, ,0 0;
−1, −1, −1, −1, −1, −1, −1, −1; −4, −2, −3, −5, −6, −3, −2, −4;
+1 (=λ).

A move (apart from castling and pawn promotion) can be specified by giving the original and final squares occupied by the moved piece. Each of these squares is a choice from 64, thus 6 binary digits each is sufficient, a total of 12 for the move. Thus the initial move P—K4 would be represented by 1, 4; 3, 4. To represent pawn promotion a set of three binary digits can be added specifying the piece that the pawn becomes. Castling is described by giving the king move (this being the only way the king can move two squares). Thus, a move is represented by (a, b, c) where a and b are squares and c specifies a piece in case of promotion.

The complete program for a type A strategy consists of nine sub-programs which we designate T_0, T_1, . . ., T_8 and a master program T_9. The basic functions of these programs are as follows:—

T_0—Makes move (a, b, c) in position P to obtain the resulting position.
T_1—Makes a list of the possible moves of a pawn at square (x, y) in position P.
T_2, . . ., T_6—Similarly for other types of pieces: knight, bishop, rook, queen and king.
T_7—Makes list of all possible moves in a given position.
T_8—Calculates the evaluating function $f(P)$ for a given position P.
T_9—Master program; performs maximizing and minimizing calculation to determine proper move.

With a given position P and a move (a, b, c) in the internal memory of the machine it can make the move and obtain the resulting position by the following program T_0.

(1) The square corresponding to number a in the position is located in the position memory.

(2) The number in this square x is extracted and replaced by 0 (empty).

(3) (*a*) If $x=1$, and the first co-ordinate of a is 6 (White pawn being promoted) or if $x=-1$, and the first co-ordinate of a is 1 (Black pawn being promoted), the number c is placed in square b (replacing whatever was there).

(*b*) If $x=6$ and a–$b=2$ (White castles, king side) 0 is placed in squares 04 and 07 and 6 and 4 in squares 06 and 05, respectively. Similarly for the cases $x=6$, b–$a=2$ (White castles, queen side) and $x=-6$, a–$b=\pm 2$ (Black castles, king or queen side).

(*c*) In all other cases, x is placed in square b.

(4) The sign of λ is changed.

For each type of piece there is a program for determining its possible moves. As a typical example the bishop program, T_3, is briefly as follows. Let (x, y) be the co-ordinates of the square occupied by the bishop.

(1) Construct $(x+1, y+1)$ and read the contents u of this square in the position P.

(2) If $u=0$ (empty) list the move (x, y), $(x+1, y+1)$ and start over with $(x+2, y+2)$ instead of $(x+1, y+1)$.

If λu is positive (own piece in the square) continue to 3.

If λu is negative (opponent's piece in the square) list the move and continue to 3.

If the square does not exist continue to 3.

(3) Construct $(x+1, y-1)$ and perform similar calculation.

(4) Similarly with $(x-1, y+1)$.

(5) Similarly with $(x-1, y-1)$.

By this program a list is constructed of the possible moves of a bishop in a given position P. Similar programs would list the moves of any other piece. There is considerable scope for opportunism in simplifying these programs; *e.g.*, the queen program, T_5, can be a combination of the bishop and rook programs, T_3 and T_4.

Using the piece programs $T_1 \ldots T_6$ and a controlling program T_7 the machine can construct a list of *all* possible moves in any given position P. The controlling program T_7 is briefly as follows (omitting details) :—

(1) Start at square 1,1 and extract contents x.

(2) If λx is positive start corresponding piece program T_x and when complete return to (1) adding 1 to square number. If λx is zero or negative, return to 1 adding 1 to square number.

(3) Test each of the listed moves for legality and discard those which are illegal. This is done by making each of the moves in the position P (by program T_8) and examining whether it leaves the king in check.

With the programs $T_0 \ldots T_7$ it is possible for the machine to play legal chess, merely making a randomly chosen legal move at each turn to move. The level of play with such a strategy is unbelievably bad.* The writer played a few games against this random strategy and was able to checkmate generally in four or five moves (by fool's mate, etc.). The following game will illustrate the utter purposelessness of random play :—

	White (Random)	Black
(1)	P–KN3	P–K4
(2)	P–Q3	B–B4
(3)	B–Q2	Q–B3
(4)	N–QB3	Q × P mate

We now return to the strategy based on an evaluation $f(P)$. The program T_8 performs the function of evaluating a position according to the agreed-upon $f(P)$. This can be done by the obvious means of scanning the squares and adding the terms involved. It is not difficult to include terms such as doubled pawns, etc.

The final master program T_9 is needed to select the move according to the maximizing and minimizing process indicated above. On the basis of one move (for each side) T_9 works as follows :—

(1) List the legal moves (by T_7) possible in the present position.

(2) Take the first in the list and make this move by T_0, giving position M_1P.

(3) List the Black moves in M_1P.

(4) Apply the first one giving $M_{11}M_1P$, and evaluate by T_8.

(5) Apply the second Black move M_{12} and evaluate.

(6) Compare, and reject the move with the smaller evaluation.

(7) Continue with the third Black move and compare with the retained value, etc.

(8) When the Black moves are exhausted, one will be retained together with its evaluation. The process is now repeated with the second White move.

(9) The final evaluations from these two computations are compared and the maximum retained.

(10) This is continued with all White moves until the best is selected (*i. e.* the one remaining after all are tried). This is the move to be made.

---

[*] Although there is a finite probability, of the order of 10^{-75}, that random play would win a game from Botvinnik. Bad as random play is, there are even worse strategies which choose moves which actually *aid* the opponent. For example, White's strategy in the following game: 1. P–KB3, P–K4. 2. P–KN4, Q–R5 mate.

These programs are, of course, highly iterative. For that reason they should not require a great deal of program memory if efficiently worked out.

The internal memory for positions and temporary results of calculations when playing three moves deep can be estimated. Three positions should probably be remembered: the initial position, the next to the last, and the last position (now being evaluated). This requires some 800 bits. Furthermore, there are five lists of moves each requiring about $30 \times 12 = 360$ bits, a total of 1800. Finally, about 200 bits would cover the selections and evaluations up to the present calculation. Thus, some 3000 bits should suffice.

6. IMPROVEMENTS IN THE STRATEGY.

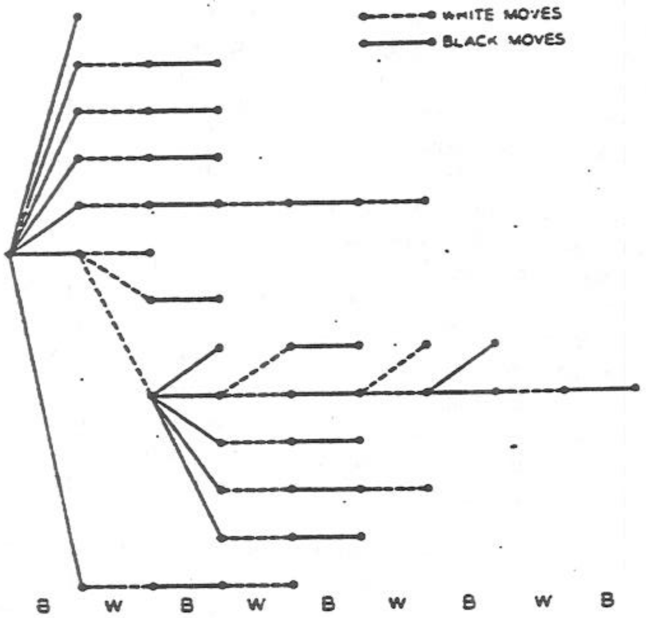
Unfortunately a machine operating according to this type A strategy would be both slow and a weak player. It would be slow since even if each position were evaluated in one microsecond (very optimistic) there are about 10^9 evaluations to be made after three moves (for each side). Thus, more than 16 minutes would be required for a move, or 10 hours for its half of a 40-move game.

It would be weak in playing skill because it is only seeing three moves deep and because we have not included any conditions about quiescent positions for evaluation. The machine is operating in an extremely inefficient fashion—it computes *all* variations to *exactly* three moves and then stops (even though it or the opponent be in check). A good human player examines only a few selected variations and carries these out to a reasonable stopping-point. A world champion can construct (at best) combinations say, 15 or 20 moves deep. Some variations given by Alekhine ("My Best Games of Chess 1924–1937") are of this length. Of course, only a few variations are explored to any such depth. In amateur play variations are seldom examined more deeply than six or eight moves, and this only when the moves are of a highly forcing nature (with very limited possible replies). More generally, when there are few threats and forceful moves, most calculations are not deeper than one or two moves, with perhaps half-a-dozen forcing variations explored to three, four or five moves.

On this point a quotation from Reuben Fine (Fine 1942), a leading American master, is interesting: "Very often people have the idea that masters foresee everything or nearly everything; that when they played P–R3 on the thirteenth move they foresaw that this would be needed to provide a loophole for the king after the complications twenty moves later, or even that when they play 1 P–K4 they do it with the idea of preventing Kt–Q4 on Black's twelfth turn, or they feel that everything is mathematically calculated down to the smirk when the Queen's Rook Pawn queens one move ahead of the opponent's King's Knight's Pawn. All this is, of course, pure fantasy. The best course to follow is to note the major consequences for two moves, but try to work out forced variations as they go."

The amount of selection exercised by chess masters in examining possible variations has been studied experimentally by De Groot (1946, *b*). He showed various typical positions to chess masters and asked them to decide on the best move, describing aloud their analyses of the positions as they thought them through. In this manner the number and depth of the variations examined could be determined. Fig. 5 shows the result of one such experiment. In this case the chess master examined sixteen variations, ranging in depth from 1/2 (one Black move) to 4–1/2 (five Black and four White) moves. The total number of positions considered was 44.

Fig. 5.



From these remarks it appears that to improve the speed and strength of play the machine must:—

(1) Examine forceful variations out as far as possible and evaluate only at reasonable positions, where some quasi-stability has been established.

(2) Select the variations to be explored by some process so that the machine does not waste its time in totally pointless variations.

A strategy with these two improvements will be called a type B strategy. It is not difficult to construct programs incorporating these features. For the first we define a function $g(P)$ of a position which determines whether approximate stability exists (no pieces *en prise*, etc.). A crude definition might be:

$$g(P)=\begin{cases}1 \text{ if any piece is attacked by a piece of lower value, or by more pieces then defences or if any check exists on a square controlled by opponent.}\\ 0 \text{ otherwise.}\end{cases}$$

Using this function, variations could be explored until $g(P)=0$, always, however, going at least two moves and never more than say, 10.

The second improvement would require a function $h(P, M)$ to decide whether a move M in position P is worth exploring. It is important that this preliminary screening should *not* eliminate moves which merely look bad at first sight, for example, a move which puts a piece *en prise*; frequently such moves are actually very strong since the piece cannot be safely taken.

"Always give check, it may be mate" is tongue-in-cheek advice given to beginners aimed at their predilection for useless checks. "Always investigate a check, it may lead to mate" is sound advice for any player. A check is the most forceful type of move. The opponent's replies are highly limited—he can never answer by counter attack, for example. This means that a variation starting with a check can be more readily calculated than any other. Similarly captures, attacks on major pieces, threats of mate, etc. limit the opponent's replies and should be calculated whether the move looks good at first sight or not. Hence, $h(P, M)$ should be given large values for all forceful moves (checks, captures and attacking moves), for developing moves, medium values for defensive moves, and low values for other moves. In explering a variation $h(P, M)$ would be calculated as the machine computes and would be used to select the variations considered. As it gets further into the variation the requirements on h are set higher so that fewer and fewer subvariations are examined. Thus, it would start considering every first move for itself, only the more forceful replies, etc. By this process its computing efficiency would be greatly improved.

It is believed that an electronic computer incorporating these two improvements in the program would play a fairly strong game, at speeds comparable to human speeds. It may be noted that a machine has several advantages over humans :—

(1) High-speed operation in individual calculations.

(2) Freedom from errors. The only errors will be due to deficiencies of the program while human players are continually guilty of very simple and obvious blunders.

(3) Freedom from laziness. It is all too easy for a human player to make instinctive moves without proper analysis of the position.

(4) Freedom from "nerves". Human players are prone to blunder due to over-confidence in "won" positions or defeatism and self-recrimination in "lost" positions.

These must be balanced against the flexibility, imagination and inductive and learning capacities of the human mind.

Incidentally, the person who designs the program can calculate the move that the machine will choose in any position, and thus in a sense can play an equally good game. In actual fact, however, the calculation would be impractical because of the time required. On a fair basis of comparison, giving the machine and the designer equal time to decide on a move, the machine might well play a stronger game.

7. VARIATIONS IN PLAY AND IN STYLE.

As described so far the machine once designed would always make the same move in the same position. If the opponent made the same moves this would always lead to the same game. It is desirable to avoid this, since if the opponent wins one game he could play the same variation and win continuously, due perhaps to some particular position arising in the variation where the machine chooses a very weak move.

One way to prevent this is to have a statistical element in the machine. Whenever there are two or more moves which are of nearly equal value according to the machine's calculations it chooses from them at random. In the same position a second time it may then choose another in the set.

The opening is another place where statistical variation can be introduced. It would seem desirable to have a number of the standard openings stored in a slow-speed memory in the machine. Perhaps a few hundred would be satisfactory. For the first few moves (until either the opponent deviates from the " book " or the end of the stored variation is reached) the machine plays by memory. This is hardly " cheating " since that is the way chess masters play the opening.

It is interesting that the " style " of play of the machine can be changed very easily by altering some of the coefficients and numerical factors involved in the evaluation function and the other programs. By placing high values on positional weaknesses, etc. a positional-type player results. By more intensive examination of forced variations it becomes a combination player. Furthermore, the strength of the play can be easily adjusted by changing the depth of calculation and by omitting or adding terms to the evaluation function.

Finally we may note that a machine of this type will play " brilliantly " up to its limits. It will readily sacrifice a queen or other piece in order to gain more material later or to give checkmate provided the completion of the combination occurs within its computing limits.

The chief weakness is that the machine will not learn by its mistakes. The only way to improve its play is by improving the program. Some thought has been given to designing a program which is self-improving but, although it appears to be possible, the methods thought of so far do not seem to be very practical. One possibility is to have a higher level program which changes the terms and coefficients involved in the evaluation function depending on the results of games the machine has played. Small variations might be introduced in these terms and the values selected to give the greatest percentage of " wins ".

8. Another Type of Strategy.

The strategies described above do not, of course, exhaust the possibilities. In fact, there are undoubtedly others which are far more efficient in the use of available computing time on the machine. Even with the improvements we have discussed the above strategy gives an impression of relying too much on " brute force " calculations rather than on logical analysis of a position. It plays something like a beginner at chess who has been told some of the principles and is possessed of tremendous energy and accuracy for calculation but has no experience with the game. A chess master, on the other hand, has available knowledge of hundreds or perhaps thousands of standard situations, stock combinations, and common manœuvres which occur time and again in the game. There are, for example, the typical sacrifices of a knight at B7 or a bishop at R7, the standard mates such as the " Philidor Legacy ", manœuvres based on pins, forks, discoveries, promotion, etc. In a given position he recognizes some similarity to a familiar situation and this directs his mental calculations along lines with greater probability of success.

There is no reason why a program based on such " type positions " could not be constructed. This would require, however, a rather formidable analysis of the game. Although there are various books analysing combination play and the middle game, they are written for human consumption, not for computing machines. It is possible to give a person one or two specific examples of a general situation and have him understand and apply the general principle involved. With a computer an exact and completely explicit characterization of the situation must be given with all limitations, special cases, etc. taken into account. We are inclined to believe, however, that if this were done a much more efficient program would result.

To program such a strategy we might suppose that any position in the machine is accompanied by a rather elaborate analysis of the tactical structure of the position suitably encoded. This analytical data will state that, for example, the Black knight at B3 is pinned by a bishop, that the White rook at K1 cannot leave the back rank because of a threatened mate on B8, that a White knight at R4 has no move, etc. ; in short, all the facts to which a chess player would ascribe importance in analysing tactical possibilities. These data would be supplied by a program and would be continually changed and kept up-to-date as the game progressed. The analytical data would be used to trigger various other programs depending on the particular nature of the position. A pinned piece should be attacked. If a rook must guard the back rank it cannot guard the pawn in front of it, etc. The machine obtains in this manner suggestions of plausible moves to investigate.

It is not being suggested that we should design the strategy in our own image. Rather it should be matched to the capacities and weaknesses of the computer. The computer is strong in speed and accuracy and weak in analytical ability and recognition. Hence, it should make more use of brutal calculations than humans, but with possible variations increasing by a factor of 10^3 every move, a little selection goes a long way toward improving blind trial and error.

ACKNOWLEDGMENT.

The writer is indebted to E. G. Andrews, L. N. Enequist and H. E. Singleton for a number of suggestions that have been incorporated in the paper.

October 8, 1948.

APPENDIX.

THE EVALUATION FUNCTION FOR CHESS.

The evaluation function $f(P)$ should take into account the " long term " advantages and disadvantages of a position, *i. e.* effects which may be expected to persist over a number of moves longer than individual variations are calculated. Thus the evaluation is mainly concerned with positional or strategic considerations rather than combinatorial or tactical ones. Of course there is no sharp line of division; many features of a position are on the borderline. It appears, however, that the following might properly be included in $f(P)$:—

(1) Material advantage (difference in total material).
(2) Pawn formation :
    (*a*) Backward, isolated and doubled pawns.
    (*b*) Relative control of centre (pawns at K4, Q4, B4).
    (*c*) Weakness of pawns near king (*e. g.* advanced KNP).
    (*d*) Pawns on opposite colour squares from bishop.
    (*e*) Passed pawns.
(3) Positions of pieces :
    (*a*) Advanced knight (at K5, Q5, B5, K6, Q6, B6), especially if protected by pawn and free from pawn attack.
    (*b*) Rook on open file, or semi-open file.
    (*c*) Rook on seventh rank.
    (*d*) Doubled rooks.
(4) Commitments, attacks and options :
    (*a*) Pieces which are required for guarding functions and, therefore, committed and with limited mobility.
    (*b*) Attacks on pieces which give one player an option of exchanging.
    (*c*) Attacks on squares adjacent to king.
    (*d*) Pins. We mean here immobilizing pins where the pinned piece is of value not greater than the pinning piece; for example, a knight pinned by a bishop.
(5) Mobility.

These factors will apply in the middle game: during the opening and end game different principles must be used. The relative values to be given each of the above quantities is open to considerable debate, and should be determined by some experimental procedure. There are also numerous other factors which may well be worth inclusion. The more violent tactical weapons, such as discovered checks, forks and pins by a piece of lower value are omitted since they are best accounted for by the examination of specific variations.

REFERENCES.

CHERNEV, 1937, *Curious Chess Facts*, The Black Knight Press.
DE GROOT, A. D., 1946a, *Het Denken van den Schaker* 17–18, Amsterdam; 1946b, *Ibid.*, Amsterdam, 207.
FINE, R., 1942, *Chess the Easy Way*, 79, David McKay.
HARDY and WRIGHT, 1938, *The Theory of Numbers*, 116, Oxford.
VON NEUMANN and MORGENSTERN, 1944, *Theory of Games*, 125, Princeton.
VIGNERON, H., 1914, *Les Automates*, La Natura.
WIENER, N., 1948, *Cybernetics*, John Wiley.

247

4-6

1

WHARTON REPROGRAPHICS

12

J. SCOTT ARMSTRONG and FRED COLLOPY*

Managers are often advised, "beat your competitors," which sometimes contrasts with the advice, "do the best for your firm." This may lead managers to focus on comparative measures such as market share. Drawing on game theory, the authors hypothesize that managers are competitor oriented under certain conditions, in particular, when they are provided with information about competitors' performance. Empirical studies lead to the additional hypothesis that a competitor orientation is detrimental to performance. To examine these hypotheses, the authors conduct two studies. The first is a laboratory study in which 1016 subjects made pricing decisions. When information about the competitor's profits was provided, over 40% of the subjects were willing to sacrifice part of their company's profits to beat or harm the competitor. Such competitor-oriented behavior occurred across a variety of treatments. The second is a field study used to examine the performance over a half-century of 20 large U.S. firms with differing objectives. Firms with competitor-oriented (market share) objectives were less profitable and less likely to survive than those whose objectives were directly oriented to profits.

Competitor Orientation: Effects of Objectives and Information on Managerial Decisions and Profitability

Traditional economic theory calls for firms to maximize shareholder wealth. However, managers often do not explicitly pursue the maximization of profits (Mueller 1992). Instead, they frequently make decisions so as to perform well relative to their competitors, which we refer to as having *competitor-oriented objectives*. Managers might choose not to focus on maximizing future profits, because profits are difficult to forecast or because a focus on profits may lead to a short-term orientation at the expense of long-term considerations. Because it is difficult to determine how close a firm is to "maximum profits," managers may worry that using such a measure could make them or their organizations content with mediocre performance.

Some experts argue that if profits do not motivate managers effectively, relative measures, such as market share. should be used. Kotler (1988, p. 333) states that the strategic objective of many firms "is to increase their marke shares, thinking that this will lead to greater profitability.' Compared with profit maximizing, such competitor-oriented objectives may be more visible because the performanc of another firm serves as a benchmark. In the face of a uncertain future, market share can serve as an intermediat goal that positions the firm for long-term profitability.

Competitor-oriented objectives are advocated in text books, magazines, and journals. For example, Hendo (1986, p. 1) stresses beating competitors at the expense c all other objectives, including profit maximization. He state that "Business is war!... It is a zero sum game.... Militar people want to destroy a target; business executives want t eliminate a competitor." Parks, Pharr, and Lockeman (199 p. 68) say, "Just as in war, the reality of market-share battl

---

*J. Scott Armstrong is Associate Professor of Marketing, the Wharton School, University of Pennsylvania. Fred Collopy is Assistant Professor of Management Information and Decision Systems, the Weatherhead School of Management, Case Western Reserve University. Robert H. Colgrove assisted in the problem formulation and data collection for the laboratory experiment. Richard H. Franke was especially helpful in guiding our thinking. Jonlee Andrews, Hector Bogo, Richard J. Boland, Gregory S. Carpenter, Douglas Cowherd, George Day, Peter R. Dickson, Jehoshua Eliashberg, Peter Fader, John U. Farley, Robert Fildes, James W. Gentry, David E. Griffith, Robin Hogarth, Graham Hooley, Raymond Hubbard, Aron Katsenelinboigen, Edwin A. Locke, Sharon McCarthy, Shelby McIntyre, Glen Omura, John Roberts, William Ross, Stephen C. Sanders, Steven P. Schnaars, Ulrike Schultze, Warren Thorngate, U. N. Umesh, Joel E. Urbany, Eric Waarts, Molly Watson, Dick R. Wittink, J. Thomas Yokum, and anonymous *JMR* reviewers provided useful comments. Jennifer L. Armstrong, Suzanne Berman, Gina Bloom, Mary Haight, Vanessa Lacoss, Martha Lightwood, and Phan Lam provided editorial assistance. The authors thank the Strategic Management Group (Philadelphia) and IDEA (Buenos Aires) for their cooperation, and the Huntsman Center at the Wharton School, which provided partial funding.

Journal of Marketing Research
Vol. XXXIII (May 1996), 188–199

is that the success of the victor depends upon the failure of the loser."

Over the past half-century, improvements in information systems have made competitor-oriented information—such as market share—more timely, accurate, detailed, and cost-effective, thus making it easier to pursue competitor-oriented objectives. One recent development is the use of scanner data. Some experts, for example, consider that the goal of using scanner data is to "take market share away from the other guy."[1]

We examine how competitor-oriented objectives and the availability of competitor-oriented information can affect managerial decisions and the profitability of firms. Using a variety of evidence collected over nine years, we compare the long-term profitability of competitor-oriented and self-oriented objectives.

Students' and Managers' Beliefs About Competitor-Oriented Objectives

Textbooks and experts advocate the use of competitor-oriented objectives—but this would not matter if their advice were ignored. To determine whether these views are held by faculty, students, and managers, we conducted a series of convenience surveys.

In surveys of marketing faculty and students in 1992, we asked: "What do you believe would be the effects on *long-term profitability* if a firm has as its primary goal to achieve higher market share?"[2] Of the 102 respondents, 57% believed that profits would be higher, 27% believed that they would be lower, and 16% were undecided.

Between 1989 and 1994, we asked 170 Masters of Business Administration (MBA) students whether "the primary purpose of the firm is (a) to do better than its competitors, or (b) to do the best that it can." One-third of these students believed that the primary purpose of the firm is to do better than its competitors. About the same number agreed with the statement that "the best way to judge the success of a firm is by how well it does relative to its competitors." In early 1995, we asked the same questions of 54 students at a university in Korea; 39% and 41%, respectively, agreed with these statements.[3]

Many managers believe that their firms use competitor-oriented objectives. In 1993, we surveyed 72 managers from a variety of firms who were attending seminars in Cleveland, Buenos Aires, and Santiago. Half of the respondents agreed that "the primary purpose of our firm is to be better than its competitors." Thirty-nine percent agreed that "the best way to judge the success of our firm is by how well it does relative to our competitors."

To examine further whether managers use competitor-oriented objectives, we examined Japan, where experts (e.g., Abegglen and Stalk 1985) claim that competitor-oriented objectives are common. Only 29% of senior executives from Japanese companies (n = 21) said that the primary purpose of their firms is to be better than their competitors, and 48% judged their success by how well they do relative to their competitors.[4]

It appears, then, that many business professors, students, and managers believe that a competitor orientation is desirable and will enhance firm performance. This suggests that there are at least two competing viewpoints about the objectives of firms. The first is that managers should pursue profits as directly as possible, and the second is that managers should use competitor-oriented objectives (such as market share) as a path to long-term profits. In the subsequent section, we discuss our hypotheses and related research.

HYPOTHESES AND RELATED RESEARCH

Studies using decisions in simple games suggest that some people adopt competitor-oriented objectives. Kuhlman and Marshello (1975) summarized research from three such studies. The share of people that selected competitor-oriented responses ranged from 21% to 49%, depending on game instructions and payoffs. Liebrand and van Run (1985) found similar results across cultures.

Griesinger and Livingston (1973) concluded from a review of experimental game literature that the proportion of "competitive" subjects varies depending on personality, cultural factors, and situational factors, such as feedback and game instructions. They also concluded (p. 187) that many subjects "seemed eager to know what was expected of them in order that they might adopt an appropriate orientation to the game." For example, the results from laboratory studies by Deutsch (1958, 1960) imply that explicit support for competitor-oriented objectives has a strong effect.

We hypothesized that under certain conditions some managers adopt competitor-oriented objectives. We further hypothesized that when competitor-oriented objectives are adopted, profits are reduced. Note that these hypotheses deal with objectives, not strategies. It seems reasonable that decision makers should consider competitors' actions when developing strategies. For example, it might be useful to examine competitors' actions for ideas about design, price, and service. Also, managers should probably consider how competitors will react to their actions.

Conditions Under Which Managers Adopt Competitor-Oriented Objectives

Social comparison theory (Suls 1977) suggests that under some conditions, people judge their performance relative to others. We expect that certain identifiable conditions might lead managers to act competitively, such as when information about competitors is made available to them. Similarly, when managers use techniques that focus attention on their company's relative performance, they are more likely to seek to maximize the competitor-oriented measure of performance than the profits.

Availability of competitor-oriented information. Dollar-auction studies demonstrate competitor-oriented behavior when one has direct knowledge about the competitor's response. Subjects bid for a dollar, with a rule that only the top bidder wins, but the top two bidders must pay. Final bids by each of the top bidders are usually well in excess of a dollar. The primary motivation of these bidders is to prevent their competitor from winning. These studies have been replicated in many countries (Teger 1980).

---

[1] Stated by James Andress, president of Information Resources in a speech at the Wharton School in 1992.

[2] Respondents were marketing faculty and students at the European Marketing Association Conference in Aarhus, Denmark (n = 13), the University of Auckland (n = 12), the University of Canterbury (n = 53), and the University of Otago (n = 24).

[3] Wujin Chu collected these survey responses.

[4] Hotaka Katahira conducted this survey.

The prisoner's dilemma has been widely used to examine decision making. Subjects can gain by defecting, but if both parties defect, they are worse off. In one such study, Corfman and Lehmann (1994) asked subjects to make "advertising spending decisions as marketing managers of a medium-sized manufacturer selling in mature markets" and to assume they were committed to remaining with the company for five years. The advertising decision involved high (competitive) or low (cooperative) budgets, and their 57 subjects chose high 78% of the time. Griffith and Rust (1994) compared the performance of subjects (MBA students) against normative pricing algorithms; their decision makers appeared to emphasize relative performance against competitors, even when they were explicitly instructed to maximize profits and their monetary compensation was based solely on profits.

One of the difficulties in assessing whether subjects are employing competitor-oriented objectives is that the desire to improve profit might explain some of the results. By revising the prisoner's dilemma payoff matrix, it is possible to remove the profit motive as an explanation. Scodel and colleagues (1959) used a matrix in which the payoff to a player does not vary when the other subject is cooperative. Each cell shows Player One's score followed by Player Two's:

	Player Two	
Player One	Black	Red
Black	3,3	1,3
Red	3,1	0,0

Assume that players were profit maximizers. Player One would select black if he expected Player Two to select red, because he would earn one point rather than zero. He would be indifferent if he thought that Player Two would select black, earning 3 points either way. The matrix is symmetrical, so neither player should prefer red in either condition. Interestingly, the proportion of red-red decisions exceeded black-black for 20 of the 22 pairs of subjects tested by Scodel and colleagues (1959) in a multiperiod experiment. In terms of profit maximizing, this behavior is irrational; it is rational, however, if the players have competitor-oriented objectives—only by selecting red does a player have a chance of doing better than the competitor.

Messick and Thorngate (1967) examined a payoff matrix in which subjects would *lose* money by making competitive decisions if the other player were cooperative:

	Player Two	
Player One	Black	Red
Black	8,8	1,2
Red	2,1	5,5

Subjects were instructed "to earn as much money for themselves as possible," and a monetary payoff, based on the number of points scored, was provided. The highest payoff comes when both play black. However, only by choosing red can one player do better than the other. When provided with feedback only on their own earnings, about one-third selected red, the competitive decision. When given feedback about other players' gains as well, most subjects selected red; on their last 20 decisions, almost 90 percent of these subjects' choices were red.

These findings about the impact of competitive information in two-player games are robust. Messick and Thorngate (1967) and Messick and McClintock (1968) examined variations of the preceding payoff matrix. They concluded that subjects primarily wanted to avoid decisions in which their earnings might be less than their competitor's. Messick and McClintock (1968, p. 23) also concluded that "subjects receiving relative-score information played to maximize the difference between their own and other player's scores more frequently than subjects given only [their] own ... scores."

In a quasi-experiment involving food stores in four U.S. cities, Boynton, Blake, and Uhl (1983) found that the publication of average food prices at various stores had little impact on consumer behavior. But it did lead stores to respond to one another such that the average prices dropped by approximately two percent. As a result, this competitive information reduced profits. Leeflang and Wittink (1996), in a study of seven large brands of a nondurable, nonfood products sold in the Netherlands, concluded that managers overreacted to each other's promotional activities. In a replication in New Zealand of the Leeflang and Wittink study, Brodie and Bonfrer (1996) found even stronger evidence that managers were too competitor oriented.

Application of competitor-oriented techniques. Many management techniques focus attention on market share. For example, portfolio planning methods examine performance relative to the competitors'. Use of these methods seems to be detrimental to profits. In laboratory experiments with 1015 subjects, Armstrong and Brodie (1994) found that use of the Boston Consulting Group matrix as a decision aid substantially reduces the profitability of subjects' decisions. Slater and Zwirlein (1992) concluded from a study of 129 firms that those whose strategies are consistent with portfolio planning models have lower returns to shareholders. Capon, Farley, and Hulbert (1987, pp. 316–17) found that firms that use the Boston Consulting Group portfolio matrix methods report a lower return on capital than those not using them.

Effects of Competitor-Oriented Objectives on Performance

It is well established that objectives affect performanc (Locke and Latham 1990). In particular, objectives have : strong effect when they are made explicit. These finding: apply to groups as well as to individuals (O'Leary-Kelly Martocchio, and Frink 1994).

When discussing the effects of a competitor orientatio on performance, people often generalize from sports, wa and other areas. Yet, their conclusions may be suspect eve in those areas. Kohn's (1986) review of competition cor tains 388 references drawn from a variety of areas, such : sports, education, and the performing arts (but not fro business). He concluded that though there may be cond tions under which competitor-oriented objectives impro performance, in general, they do not.

Campbell and Furrer (1995), in an experiment involvi the solution of routine arithmetic problems, concluded th competition had "a significant dysfunctional effect on task performance." In Deutsch's (1958, 1960) studies, when participants were told to do better than their opponents, both parties ended up worse off than when told to do the best they could for themselves.

Market share is often used as a relative measure of performance, and it may be a surrogate for the measure of true interest, namely, long-term profit. Because it does not directly assess what is of interest, there is a potential for other factors to intervene, causing it to give a less accurate representation of the effects on profits. Moreover, basing decisions on attaining market share can have harmful effects, such as price wars. Cassady, who examined actual price wars, concluded that, in some situations, "vendors temporarily shift their emphasis away from attaining success [for themselves] and toward preventing the success of rivals" (summarized by Teger 1980). Similarly, Anterasian and Graham (1989) found that detrimental performance resulted from competitor-oriented objectives. They used a simulation of firms to compare a self-oriented goal (stability) against a competitor-oriented goal (market share). The market share goal reduced profits when the market had cycles.

Szymanski, Bharadwaj, and Varadarajan (1993) conducted a metanalysis of studies relating market share and profitability. These cross-sectional studies showed positive correlations between market share and profitability. Many observers have concluded that this relationship is causal, though others (e.g., Jacobson and Aaker 1985; Prescott, Kohli and Venkatraman 1986) disagree. For example, firms with profits as their sole objective might produce superior products and, as a result, achieve gains in both market share and profits. Moreover, and of critical importance to the issues raised here, these studies did not examine the *objectives* of the managers.

Studies that have used a longitudinal, rather than cross-sectional, approach have found a negative relationship between market share and profits. Anterasian and Graham (1989) analyzed data on 42 firms in industries that had cycles; companies that lost market share during growth periods tended to be more profitable over the cycle than firms in the same industry that gained market share. Tschoegl and Yu (1990), in a study of the liquor market, found that a high market share did not help in gaining further share and did not produce stability in the firm's sales. Montgomery and Wernerfelt (1991) examined the performance of six large U.S. brewers from 1969 to 1979, a period characterized by large changes in market share; using returns on stocks, they concluded (p. 958) that gains in market share were associated with "the destruction, rather than the creation, of firm value."

LABORATORY EXPERIMENTS

We conducted laboratory experiments to examine whether managers make competitor-oriented decisions even when they know in advance that such decisions are harmful to profits. The decisions of subjects who were given information only on their own firm's profits were compared with those of subjects who also were given information on their competitor's profits.

Administration

In the subsequent discussion, a "treatment" refers to the description provided to the subjects (e.g., "harm" or "beat"). These treatments were used to examine alternative explanations. Treatments were randomly assigned, and each subject received only one. An "administration" refers to the treatments tested at a given time. Some administrations had few subjects, so few treatments were used. After making their decisions, subjects responded to written follow-up questions about reasons for their decisions. Most administrations were concluded with a debriefing of the subjects.

Subjects

Managers or potential managers were sought as subjects. The subjects (n = 1016) came from two major business schools in the United States (73 undergraduates, 846 MBAs, 42 executive MBAs) and from a business school in Argentina (20 MBAs, 35 executive MBAs).

The subjects were familiar with the type of problem presented, through either academic training or work experience. Almost all had full-time work experience, and their managerial experience was extensive. The experiments were conducted in classroom settings, and the subjects worked individually. This use of captive subjects reduced self-selection bias.

Materials

Subjects were asked to assume the role of a marketing manager for a company called Big Guys, Inc. (To avoid associations with actual firms, names that sounded obviously fictitious were used.) Subjects were told, "We are performing an experiment in decision-making and we need your help.... You will only need about 10 or 15 minutes to complete the experiment."

The subjects received written descriptions of a business situation that was more extreme than might be encountered in the real world. The decision involved a choice of a "high" or "low" price for a new product. Five-year profit forecasts were provided for each pricing decision. Our intent was to convey uncertainty about profits after the stated planning horizon. A base treatment made no mention of competitors:

> You are the marketing manager of a manufacturing firm known as Big Guys, Inc. As the company's marketing manager, you are responsible for all marketing decisions and strategies, including the pricing structure of the firm's products. Recently your company introduced a new highly technical product, and you have been asked to set the pricing strategy for this product. You calculate the present value of the total profits expected for your firm over the next five years. You determine the following results for both strategies:

Expected Profits Over Five Years

Low-Price Strategy	*High-Price Strategy*
$40 million	$80 million

The harm and beat treatments added information about the impact of the subjects' decisions on competitors.

Harm treatment. In this treatment, subjects were able to harm their competitors. Subjects were told that their main competitor had a similar product and that the market for both products was the same. The following was added to their description:

> You are aware that your main competitor, Other Guys, Inc., intends to introduce a product that is *very* similar to the one that your firm has just introduced. You should assume that the competitor's product is as good as yours in every way that is important to the market, and the market is the same for both products. Therefore, the pricing strategy which you must formulate for your product should take into account this competitive force.
>
> You are essentially faced with two choices:
>
> 1. Keep your price low, which causes your competition to suffer a substantial loss, or
> 2. Choose a higher price that produces higher profits for your firm, but which also allows the competition to prosper.
>
> You then calculate the present value of the total profits expected for your firm over the next five years, as well as for the competitors (Other Guys). You note that after five years the products are in the mature stage of the product life cycle. You determine the following results for the two strategies:

Expected Profits Over Five Years

	Low-Price Strategy	*High-Price Strategy*
Big Guys	$ 40 million	$80 million
Other Guys	−100 million	40 million

Subjects could earn $140 million more than competitors by choosing the low price, but only $40 million more by choosing the high price. In terms of *relative gains*, then, the low price is best.

Beat treatment. Because the harm treatment inflicts heavy losses, subjects might infer that this leads to the demise of the competitor and discourages future entry. If some subjects were simply making judgments they believed would lead to long-term profits, then the elimination of this damage to the competitor might reduce the number making this choice. The beat treatment addresses this.

The description was similar to the harm treatment. Outcomes for the subject's firm were the same as in the harm treatment, but the competitor's profits were $120 million higher for each alternative. At the low price, the subjects' firm earns $20 million more in profits than the competitor, but the competitor apparently operates successfully. Profits are higher for the high-price decision, but here the competitor earns more than the subject's firm. Thus, only by choosing the lower price can the subject do better than the competitor, whereas in the harm treatment the firm would "win" with either strategy.

Expected Profits Over Five Years

	Low-Price Strategy	*High-Price Strategy*
Big Guys	$40 million	$ 80 million
Other Guys	20 million	160 million

Face Validity of the Instruments

To assess whether the situations allowed for profit maximization, they were presented to economics and finance professors at a major university suggested by the chair of the economics department on the basis of relevant experience. Each faculty member received the harm and beat treatments and was asked to assume that he (all faculty members were male) had been called in as an economic advisor to the firm. Six faculty members responded, and all agreed that the high price was the profit-maximizing decision for each treatment; when asked, they expressed no need for further information.

Base Results

When no information was provided about the performance of competitors, 14% of 65 subjects selected the less profitable decision. On the follow-up questionnaire, some subjects who had selected the low price expressed concern about their competitors.

In the harm treatment, 34% of the 139 subjects selected the less profitable decision. Thus, here subjects were twice as likely to select the less profitable decision as when no information was provided about competitors ($Z = 2.66$; $p < .01$).

In the beat treatment, 60% of the 60 subjects selected the low price. Testing against the null hypothesis that the information about competitor's performance had no effect (the beat treatment versus the base treatment with no information about competitors) produced a significant rejection ($Z = 5.86$; $p < .01$).

Replications with Extensions

To assess reliability and address threats to validity, 23 different treatments using 43 experimental administrations were conducted from 1986 to early 1995. One concern in these extensions was to assess the alternative explanation that subjects selected the low-price decision so as to maximize long-term profits. These extensions involved treatments that lengthened the time horizon, changed the situation from the future to the past, and equalized the final market values.[5]

Lengthening time horizon. Because subjects might be willing to suffer short-term losses in exchange for long-term gains, the time horizon was lengthened from 5 to 20 years for the harm and beat treatments.[6] Subjects were told that the matrix represented net present values of profits over a 20-year period. For the 42 subjects who did not receive information about the performance of their competitor in the pricing treatment, 12% made the less profitable decision. In contrast, 45% of the 40 harm treatment subjects and 30% of the 40 beat treatment subjects selected the less profitable decision. The difference between the subjects with and without information on competitors was significant for both the harm ($Z = 3.02$; $p < .01$) and beat treatments ($Z = 1.65$; $p < .05$). Lengthening the horizon had little effect on results in the harm treatment, but it reduced the percentage choosing the low-price strategy in the beat treatment; we have no explanation for this result.

---

[5] Minor extensions examined the nature of the product. The problem was initially described as involving a "new, highly technical product." To assess whether the type of product made a difference, the product was varied in treatments that did not mention competitors. One treatment, administered to 15 subjects, specified no product; here, only one subject chose the less profitable decision. Another treatment said that the product was "a new perfume for women"; three of 17 subjects chose the less profitable decision. For these two treatments, then, 12.5% of the 32 subjects chose the less profitable product. This is similar to the 13.3% of the 75 subjects who had received the technical product treatment. We concluded that the type of product did not have an important effect and have combined all these results in the analysis of the base treatment.

[6] This change was also made for the ex post pricing decisions that are described in the next section.

Ex post pricing decision. Despite the instructions, subjects might have made different assumptions about profits beyond the stated planning horizon. To control for this, the problem was changed to refer to the past. Here, subjects had to decide which of two brand managers should be promoted by the company, called Multiple Products, Inc. In this "ex post pricing decision," the choices were between a manager who used a high-price strategy and one who used a low-price strategy. Both had managed in comparable but independent geographical regions for the same five years. The profit matrix was identical with that in the original harm pricing decision: The manager choosing the high price earned high profits and his competitor also earned high profits, whereas the manager choosing the low price earned lower profits and his competitor's profits were negative. The description stated that "whatever their profit situation has been over the past five years, all of the competitors still exist in their respective regions."

Of the 76 subjects in the ex post harm pricing treatment, 40% promoted the marketing manager who used the low price. Of the 127 subjects in these same administrations who had received the basic harm pricing treatment, 34% selected the less profitable decision. Thus, these results are not consistent with a hypothesis that the subjects were trying to increase long-term profits.

Similarly, of the 69 subjects in an ex post beat treatment, 51% promoted the manager who used the low-price strategy. This is comparable to the 60% of the subjects who selected the low-profit decision in the basic beat treatment in the same administrations.

Equalizing final market values. A further revision of the ex post pricing treatments added information that the final market values of the company's two divisions were the same:

> Currently, the firm has a cash flow shortage. Over the past seven years the firm has maintained its dividend payout in the face of declining profits, such that all profits have been paid out. As a result, the firm has decided to sell one, but not both, divisions. In trying to decide which division to sell, both were put on the market. Bidding has been intense and the selling price for each division is almost identical. These have been good divisions for the company. They had been acquired ten years ago and the purchase price for each was 40% of the current market price. Marketing for the two divisions are headed by Harry Jones and Louie Smith. You must decide whether to retain Harry or Louie for the division that the company will keep.

In other words, the two divisions were valued equally initially and also currently, but the manager who used the high price achieved much higher profits during the intervening period.

Of the 87 subjects receiving the equal–market-value harm treatment, 52% promoted the manager who had the lower profits. Surprisingly, this exceeded the 34% of such promotions by the 80 subjects in the equal–market-value beat treatment. Given complete information about profitability, then, 43% of the 167 subjects selected the competitor-oriented (less profitable) option.

Exposure to Competitor-Oriented Techniques

We anticipated that competitiveness would intensify when people were exposed to techniques that advocate competitor-oriented objectives. To test this, some subjects were provided with a description of the experience curve, which states that reductions in unit costs occur as cumulative volume increases. The experience curve strategy advises firms to cut prices to build volume, in order to propel the firm down its cost curve faster than competitors can move down theirs. In other words, they should price so as to prevent competitors from catching up. Lieberman (1987, p. 451) concluded that the experience curve produces incentives that "often intensify competition and reduce profits."

To assess the impact of exposure to the experience curve, some subjects received a description from Kiechel (1981, pp. 139–140) advocating its use. The decisions of those 97 subjects were compared with those from 137 control subjects in the same administrations; the latter received no information about the experience curve. More experience curve subjects selected the less profitable decision than did control subjects (59% versus 45%; $p < .05$ using chi-square).

A neutral description of the experience curve from Porter (1980) was provided to another set of subjects. Of the 63 subjects receiving this description, 41% selected the less profitable decision, and 42% of the 86 control subjects selected it. By this test, the neutral description of the experience curve had no significant effect.

In a before-and-after design, 113 subjects made decisions before reading the neutral description of the experience curve. Then they read it and made a second decision. Nine subjects changed from a high to a low price decision, and none went in the other direction. The percentage who chose the lower-profit choice went from 43% to 51%. This was a modest but statistically significant increase ($p < .05$ based on the McNemar change test with correction for continuity from Siegel and Castellan 1988).

Overall, then, information about the experience curve produced less profitable decisions for two tests, with no difference on one. This provides moderate support for our hypothesis.

Management Training

Management training often emphasizes competing. For subjects from one of the U.S. business schools, decisions made by those who had more formal management education were compared with decisions by subjects who had less. Subjects just entering the management program were placed in a "low" group, those who had had courses in strategic planning were put in a "high" group, and the rest were placed in an "intermediate" group. This classification was rough, because subjects within each group differed with respect to their prior management education and the treatments were not balanced across education levels. Furthermore, the competitor orientation was probably high among most subjects prior to this management program.

These shortcomings should reduce the estimated effect of education on decisions.

Using decisions by subjects in the harm and beat treatments, education levels were compared. Of the 236 subjects in the low education group, 38% selected the less profitable decision. In contrast, 46% of the 227 subjects in the intermediate education group and 55% of the 88 subjects in the high education group chose the less profitable decision. The differences among groups were statistically significant ($p < .05$ using chi-square). In effect then, those with more management education made less profitable decisions.

Follow-Up Questionnaire

In follow-up questionnaires from 14 administrations of the experiment, subjects were asked, "Why did you make this decision?" In general, once subjects took the competitor's performance into account, they were less likely to pay attention to their own profits. Eighty-three percent of the 107 subjects who mentioned competitors selected the less profitable decision.

Sixty-seven percent of the 133 subjects who selected the low-price decision made statements about beating their competitors, while only 38% discussed profits. Some of those selecting the low price said that the company's objective was to increase market share, that it was important to beat competitors, or that they wanted to be the leader. Some respondents commented on their willingness to sacrifice profits. One respondent who had been employed by an international consulting firm stated, "Perhaps it's my strategic planning background, but I would definitely pay to harm my competitors." Another subject who selected the less profitable decision said, "Louie's strategy [low price, competitor-oriented] has generated twice the profits of the competitor, while Harry's strategy [high price] has reaped half the profits of the competitor. Hence, Louie's strategy seems to be the strongest. Destroy the competitor."

Of the 208 subjects selecting the high-price decision, 78% explicitly discussed profits, while fewer than 10% mentioned competitors. Of the latter, most said that they did not want to harm their competitor. Some said that they were pleased that their competitor would also succeed ("a win-win situation"). Others stated that the competitor's performance was not relevant to their decision.

Limitations and Threats to Validity of the Laboratory Studies

The use of students as subjects posed few threats to the studies' validity. Prior research, summarized by Ashton and Kramer (1980), found similarity between students and nonstudents in studies of *decision making*, which is the concern of our study. Also, consistent with criteria suggested in Gordon, Slade, and Schmitt (1987), our subjects were similar to the population of interest. Corfman and Lehmann (1994) found no substantial differences between executives and students in their prisoner's dilemma study of advertising decisions.

Table 1

PERCENTAGE OF SUBJECTS WHO SELECTED THE LESS PROFITABLE DECISION (NUMBER OF SUBJECTS)

	Information About Competitors		
Treatment	*None*	*Harm*	*Beat*
Pricing:			
5 years (benchmark)	14 (65)	34 (139)**	60 (60)**
20 years	12 (42)	45 (40)**	30 (40)*
Unweighted Averages	13	40	45
Ex Post Pricing:			
5 years	n.a.†	40 (76)	51 (69)
20 years	n.a.	25 (24)	30 (30)
5 years: equal value	n.a.	52 (87)	34 (80)
Unweighted Averages		39	38

*Significant at $p < .05$ with one-tail t-tests (compared with results from "none" column).

**Significant at $p < .01$.

†No tests could be conducted for the ex post treatments, because we could not construct a "no information" version.

Results from the experience curve treatment might be due to demand effects. Providing information about the experience curve might have led subjects to assume that it was a relevant technique, but it is difficult to use demand effects to explain the differences in decisions across levels of education. In the other parts of the experiments, we had no reason to expect demand effects. We did consider the possibility that subjects might answer so as to look good to the experimenter (Sigall, Aronson, and van House 1970). To do this successfully, they would have had to be aware of the purpose of the study. Follow-up discussions with subjects revealed some awareness. When 84 of them were asked to guess the purpose, ten (12%) guessed correctly. Only a small percentage, then, would have known how to modify their responses to look good in this situation. Nevertheless, this does not rule out demand effects (Shimp, Hyatt, and Snyder 1991).

Although researchers in management have become increasingly critical of the external validity of laboratory studies (Griffin and Kacmar 1991), evidence suggests that their validity is comparable to that of field experiments (Locke 1986).

Summary of Laboratory Results

The results, summarized in Table 1, were consistent with our hypotheses, the effects were strong, and they were statistically significant: A substantial portion of the subjects selected competitor-oriented decisions. This provides evidence of economically irrational behavior.

FIELD STUDY

Historical performance of American industrial firms between 1938 and 1994 was used to determine the extent to which a competitor orientation can affect profitability. To assess what marketing experts would expect from such a study, we posed the following question to convenience samples of 90 marketing faculty members in New Zealand and the United States, as well as to 18 managers in Buenos Aires:

> What if we ran the following study: (1) select 20 firms from different industries; (2) assess the extent to which

> their goals are competitor oriented (market share); and (3) examine their profits over the next three decades. Assuming that the 20 firms would differ greatly with respect to competitor orientation, what would you predict?

The key item was: "Profits in firms with market share as a primary goal are (1) much less, (2) less, (3) the same, (4) more, or (5) much more than in firms with profit-oriented goals." Of the 108 respondents, 52% said that profits would be more or much more, and 26% thought they would be less or much less.

Information about firms' competitor orientations and their long-term profitability was needed to test the relationship. Fortunately, information on pricing objectives of 20 large U.S. companies had been collected by Lanzillotti (1958), who, sponsored by the Brookings Institution, investigated the price-setting process. He examined "the motivational hypothesis of the firm, i.e., the specific objectives upon which business firms base pricing decisions."

Competitor Orientation of 20 Firms

The 20 corporations, drawn from a variety of industries, were among the 200 largest U.S. industrial corporations in terms of assets. According to Lanzillotti (1958, p. 921–22), "The companies were selected from among the largest corporations on the basis of the willingness of management to cooperate by permitting extensive interviews with top company officials," and because "the prominence of each of the corporations in their respective industries made them masters, to a significant degree, of their fates; hence they were able to adjust pricing to the company's general goal."

The competitor orientation of the firms' objectives was inferred from Lanzillotti's description of their stated pricing policies. Between 1948 and 1951, he conducted interviews with company officials that lasted approximately one week in each company. Lanzillotti (1958) explained that such a careful procedure was needed because of the sensitive nature of pricing policies. A second set of interviews was carried out in 1956–1957 to fill gaps in the study and determine whether changes had occurred in pricing policies since the original interviews (Kaplan, Dirlam, and Lanzillotti 1958). Judging from Kaplan, Dirlam, and Lanzillotti's (p. 187) study, the competitor orientations were stable from the initial interviews to the follow-up interviews, a period of nine years. They also said that, in some firms, objectives were based on long-standing policies. For example, as far back as 1937, A&P had stated that their primary aim in pricing was to achieve a larger market share.

An 11-point scale (1 = to do well for themselves; 11 = to do well relative to competitors) was developed to assess firms' objectives. The objective of doing well for a person's own firm was identified by the presence of an explicitly stated pricing goal to maximize or increase profits. When profit targets were stated, a distinction was made between high and modest objectives. A priori, we classified a target as "high" if it called for maximizing profits or if it had a target return on investment (ROI) of at least 15% after taxes. At the other end of the scale were competitor-oriented objectives, stated in terms of market share. Goals of increasing or maximizing market share were classified as highly competitor oriented, and goals of maintaining market share were moderately competitor oriented. Upper limits on market share (e.g., to stay under 20%) were not coded as competitor oriented. Lanzillotti (1958) described "primary pricing goals," which were used to make basic category assignments, and "collateral pricing goals" (see Table 2). One firm did not specify objectives in terms of market share or profits, but focused on stability of operations. To handle this, an additional category was created between goals that focused only on profit and those that focused on market share.

Table 2

OBJECTIVES SCALE AND NUMBER OF FIRMS AT EACH LEVEL

Competitor Orientation 1 = low 11 = high	*Pricing Objectives*		*Number of Firms*
	Primary	*Secondary*	
1	High Profit*	n.s.†	3
2	Profit	n.s.	1
3	Stability	n.s.	1
4	High Profit	Maintain MS	2
5	High Profit	Increase MS	0
6	Profit	Maintain MS	4
7	Profit	Increase MS	0
8	Maintain MS	Profit	2
9	Increase MS	Profit	1
10	Maintain MS	n.s.	4
11	Increase MS	n.s.	2

*High profit means a return of investment of at least 15% after taxes; MS = market share.

†n.s. = "none stated."

Firms' orientations were coded by the first author and an expert on competitor-oriented behavior (Peter S. Fader). The latter received Lanzillotti's article (from which information about profitability and market share had been deleted) along with a copy of the rating scales. Intercoder reliability was high (r = .96); minor differences occurred for orientations of only four firms. Differences resulted from failures to describe how to handle an upper limit on market share and from the interpretation of Lanzillotti's description. After discussion, but before analyzing relationships, each coder changed two items; this produced full agreement.

Firms were located at all levels of the scale except 5 and 7. The median firm's orientation was 6, the scale's midpoint. Competitor-oriented objectives were adapted by 30%, and an additional 45% used a combination of profit- and competitor-oriented objectives.

Performance of the 20 Firms

Our hypothesis states that a competitor orientation harms the firm's profit; it makes no prediction about *relative* performance. Thus, the company's profits were examined in absolute terms, not in comparison with others in the same industry. Specifically, the average after-tax ROI was used. Lanzillotti (1958) provided these data over nine years (1947–1955) because, in his opinion, they measured performance that might have been affected by the pricing policies.

Return on investment was "net income to surplus" (after taxes) divided by the firm's "total assets." Restated earnings were used when available. To guard against errors, two research assistants independently collected the data, and the

Table 3

COMPETITOR ORIENTATION OF FIRMS AND ROI FOR NINE-YEAR PERIODS

Firm	*Competitor Orientation* *1 = low; 11 = high*	*ROI (After Taxes)* *1938–1946*	*1947–1955*	*1956–1964*	*1965–1973*	*1974–1982*
DuPont	1	9.1	15.4	15.5	8.0	6.9
General Electric	1	8.1	10.9	9.4	6.7	7.9
Union Carbide	1	9.8	11.0	9.1	6.3	6.6
Alcoa	2	8.5	6.4	4.2	4.2	5.5
Kennecott	3	8.6	13.3	8.9	8.2	3.2[a]
General Motors	4	8.8	16.6	13.2	12.0	6.3
Johns Manville	4	6.8	11.2	4.6	7.6	4.9[b]
Standard Oil of New Jersey (Exxon[c])	6	5.4	13.0	7.8	7.6	8.0
General Foods	6	11.9	8.2	11.4	8.9	7.4
US Steel (USX[d])	6	3.4	6.5	6.0	3.5	3.4
International Harvester[e]	6	4.7	6.7	4.6	4.0	–3.4
Kroger	8	7.2	8.0	6.1	4.9	4.6
Standard Oil of Indiana[f]	8	5.3	7.1	5.4	6.4	8.3
Sears	9	8.8	12.4	8.5	6.4	4.2
Goodyear	10	5.8	6.4	7.0	5.7	4.0
Gulf[g]	10	5.0	9.7	8.9	7.1	6.3
American Can[h]	10	6.8	7.8	5.2	4.8	3.8
Swift	10	3.9	4.6	2.4	3.3	n.a.[i]
Great Atlantic & Pacific	11	6.8	8.4	7.8	4.2	–2.9
National Steel	11	5.2	9.6	6.0	5.1	1.1[j]
Correlation With Competitor Orientation		–.54*	–.43*	–.37*	–.43*	–.45*

*Correlations were significantly different from $r = 0$ at $p < .05$.
[a] Acquired by Standard Oil of Ohio in July 1981 (1981–82 excluded).
[b] Succeeded by Manville Corporation on October 30, 1981, when the firm filed for Chapter 11 (1982 excluded).
[c] Name changed in 1973.
[d] Name changed in 1986.
[e] Name changed to Navistar in 1986 and most assets were sold.
[f] Name changed to Amoco in 1985.
[g] Acquired by Chevron in 1984.
[h] Became part of Primerica, an insurance and financial firm, in 1987.
[i] Reorganized April 30, 1973, after which Swift consisted only of food divisions (1973 and subsequent years excluded).
[j] Reorganized as National Intergroup in 1982, and the steel operations were sold to Japan's NKK Corporation.

first author checked a sample of firms.[7] (Our figures for ROI for 1947–1955, obtained from *Moody's Industrial Manual*, were somewhat different from those reported by Lanzillotti and we were unable to identify why.[8])

Competitor-oriented objectives and ROIs were negatively correlated for the 1947–1955 data. The Spearman correlation (all reported correlations were corrected for ties) was $r_s = -.43$, which was statistically different from the null hypothesis that competitor-oriented objectives are not related to the level of profitability ($p = .03$ using a one-tail test). Because many other factors also influence a firm's ROI, these results suggest a strong relationship.

This analysis of profitability covered the period coincident with Lanzillotti's assessment of pricing goals, and we also analyzed the nine years after this. In Table 3, we present the ROI for each firm for 1956–1964. The Spearman correlation between competitor-orientation and ROI remained negative ($r_s = -.37$) and statistically significant ($p = .05$)—that is, firms that had competitor-oriented objectives in the 1950s continued to earn less on their investments.

Advocates of competitor-oriented objectives suggest that gaining market share improves *long-term* profitability, so we analyzed additional nine-year periods (see Table 3). The ROIs for 1965–1973 were based on *Moody's Industrial Manual* and those for 1974–1982 on *Compustat*.[9] As before, competitor-oriented objectives were inversely related to ROI—Spearman correlations were –.43 and –.45 respectively (both significant at $p < .05$).

We also looked at survival rates of the 20 companies. *Survival* was defined here as avoiding bankruptcy or sale of the firm because of poor performance. Companies whose only goal was profit (those coded as 1 or 2 in Table 3) were compared against those with market share as their only goal (codings of 10 or 11). Information from *Moody's Industrial*

---

[7] Jennifer L. Armstrong, Michael Halperin, and Gina Bloom collected financial data for the field study.

[8] Using Lanzillotti's data, the results would have been even stronger.

[9] *Compustat* was used because it was available for 1974–1982, and it adjusted these data to aid comparability across companies, thereby aiding potential replication. The figures for *Moody's* and *Compustat* were identical for about two-thirds of the entries for 1974–1982, and the differences were typically minor for the remaining one-third. When the 1974–1984 *Moody's* data were substituted, a Spearman correlation with competitor orientation of –.36 (significant at $p = .06$, one-tail test) was obtained.

Manual through 1994 was examined, a span of 36 years since Lanzillotti's paper. All four profit-oriented firms survived, while four of the six competitor-oriented companies failed.[10] Thus, competitor-oriented firms were less likely to survive ($p = .07$ by the Fisher Exact Test).

Limitations of the Field Study

A person might argue with Lanzillotti's statement that these firms were masters of their fate, suggesting that short-term conditions within the industry led firms to adopt competitor-oriented objectives. In particular, managers in firms that were doing poorly in the mid-1950s might have decided at that time that they could improve by focusing on market share. To test this, we examined the *prior* nine-year period (1938–1946; see Table 3). If a change in performance caused the change in orientation, there should be little correlation between competitor orientation and ROI in the earlier period. As it turned out, competitor orientation was negatively related to ROI (Spearman correlation = –.54; $p < .02$).

Our analysis covered 54 years, and many changes occurred in the industries as well as in the firms. Of particular concern is that objectives of some firms might have changed. Such changes would be expected to reduce the relationship that we hypothesized.[11] Another change is that some of the firms, such as Alcoa and DuPont, have decentralized and allowed the strategic business units to set their own objectives. Finally, differences among firms might have been caused by some other factor that happened to be related to the selection of this group of firms, but we were unable to think of such a factor.

DISCUSSION

Our research strategy was one of triangulation: We proposed two hypotheses supported by prior research and then conducted a series of laboratory experiments and a field study.[12] Several different measures and procedures were used, and these were established prior to analyzing the data. The studies utilized evidence from different cultures and different time periods. This multimethod, multimeasure strategy, articulated by Campbell and Fiske (1959), has been recommended where experiments with randomly selected subjects are difficult.

What we find most striking, then, is the pattern of results. Although it is possible to provide alternative explanations for particular results, it is difficult to find an alternative that is consistent with the prior research and all our results.

On the other hand, various limitations have been identified in the previous research and our studies. Also, our results lack face validity among many experts. In addition, there may be conditions under which competitor-oriented objectives are useful. That noted, Slater and Narver (1994) examined whether the competitive environment might have an impact on the effectiveness of different corporate objectives, but they found none.

Further studies should examine alternative explanations for our findings, limits to the generalizability of our conclusions, why competitor-oriented objectives are adopted, and the impact of competitor-oriented objectives on other stakeholders. Some experts, for example, suggest that customer satisfaction should be the focus of objective setting. We have not examined this, but it could be an area that warrants research.

SUMMARY AND CONCLUSIONS

In the laboratory study, information about their competitor's performance led more than 40% of subjects to select less-profitable alternatives. Replications and extensions were conducted so that 23 treatments were conducted in 43 experimental administrations with 1016 subjects over a nine-year period. Even when net present value of profits were reported for 20 years, more than one-third of the subjects chose the less profitable alternative. Another extension cast the problem in the past by asking which of two managers should be retained by the firm; almost 40% of the subjects decided to keep the manager that sacrificed his own firm's profits to do well relative to competitors. In an extension of this ex post pricing decision, which stated that the market values of the divisions were currently equal (such that the only thing that differed was the profit that had been earned by each manager), 43% of the subjects favored the manager who produced lower profits. A competitor-oriented technique, the experience curve, led to less profitable decisions. Subjects with more management education (which presumably emphasizes competitor-oriented objectives) made less profitable decisions.

In the field study of 20 U.S. firms, those with competitor-oriented goals had lower ROIs during 1947–1955. Similar results were observed for each of three successive nine-year periods through 1982, as well as for the nine-year period preceding 1947. Firms with competitor-oriented objectives were also less likely to survive.

Our results suggest that the use of competitor-oriented objectives is detrimental to profitability. Because of this pattern of evidence, we suggest that firms should ignore their competitors when setting objectives and, instead, focus directly on profit maximization. This conclusion has implications for managers and management professors. We recommend the following:

- Do not use market share as an objective.
- Avoid using sports and military analogies, because they foster a competitor orientation.

---

[10] Each of the failed firms experienced financial difficulties. As a result, Gulf was purchased by Chevron in 1985, American Can became Primerica in 1987 and abandoned the canning business, Swift reorganized with part becoming Esmark in 1973, and National Steel reorganized in 1983 under National Intergroup (a pharmaceutical company) and was bought by Japan's NKK Corporation.

[11] For example, in December 1981, the Chief Executive Officer of General Electric (GE) publicly stated that GE's goal was to be number one or number two in each of its businesses (Slater 1993, pp. 73–76). This contrasts with GE's profit orientation in our analysis. It should be noted that GE also made other significant changes (e.g., downsizing and decentralization). The ROI in the decade prior to these changes was 7.5%, whereas in the decade after it was 6.0%. Many other factors varied over these two decades.

[12] We report on all but one of the studies that we undertook on this project. That study involved a comparison of managers' competitor orientation and the average market value of firms in a seven-year management game simulation. The results were consistent with our hypotheses ($p < .10$), but the test involved only four industries. Copies of these results are available to interested readers.

- If you use benchmarking, ensure that it does not influence objective setting.
- Do not use management science techniques that are oriented to maximizing market share, such as portfolio planning matrices and the experience curve.
- Design information systems to focus attention on the firm's performance, as measured by profits.

Because of the evidence to date, we believe that microeconomic theory, with its emphasis on profit maximization, is the most sensible course of action for firms; that is, managers should focus directly on profits. Our findings also raise concerns about information systems—having information about a competitor's performance may be harmful to a person's own performance. Improvements in the ability to measure market share, first through surveys and more recently through scanner data, might lead to a stronger focus on competitiveness and to less profitability.

REFERENCES

Abegglen, James C. and G. Stalk, Jr. (1985), *Kaisha: The Japanese Corporation*. New York: Basic Books.

Anterasian, Cathy and J. L. Graham (1989), "When It's Good Management to Sacrifice Market Share," *Journal of Business Research*, 19 (3), 187–213.

Armstrong, J. Scott and R. J. Brodie (1994), "Effects of Portfolio Planning Methods on Decision Making: Experimental Results," *International Journal of Research in Marketing*, 11 (1), 73–84.

Ashton, R. H. and S. S. Kramer (1980), "Students as Surrogates in Behavioral Accounting Research: Some Evidence," *Journal of Accounting Research*, 18 (1), 1–16.

Boynton, Robert D., B. F. Blake, and J. N. Uhl (1983), "Retail Price Reporting Effects in Local Food Markets," *American Journal of Agricultural Economics*, 65 (1), 20–29.

Brodie, Roderick J. and A. Bonfrer (1996), "Do Managers Overreact to Each Others' Promotional Activity? Further Empirical Results," *International Journal of Research in Marketing* (forthcoming).

Campbell, Donald J. and D. W. Furrer (1995), "Goal Setting and Competition as Determinants of Task Performance," *Journal of Organizational Behavior*, 16 (4), 377–89.

Campbell, Donald T. and D. W. Fiske (1959), "Convergent and Discriminant Validation of the Multitrait-Multimethod Matrix," *Psychological Bulletin*, 56 (2), 81–105.

Capon, N., J.U. Farley, and J. M. Hulbert (1987), *Corporate Strategic Planning*. New York: Columbia University Press.

Corfman, Kim P. and D. Lehmann (1994), "The Prisoner's Dilemma and the Role of Information in Setting Advertising Budgets," *Journal of Advertising*, 23 (2), 35–48.

Deutsch, Morton (1958), "Trust and Suspicion," *Journal of Conflict Resolution*, 2 (4), 265–279.

——— (1960), "The Effect of Threat Upon Interpersonal Bargaining," *Journal of Abnormal and Social Psychology*, 61 (2), 181–89.

Gordon, Michael L., Allen Slade, and N. Schmitt (1987), "Student Guinea Pigs: Porcine Predictors and Particularistic Phenomena," *Academy of Management Review*, 12 (1), 160–63.

Griesinger, Donald W. and J. W. Livingston, Jr. (1973), "Toward a Model of Interpersonal Motivation in Experimental Games," *Behavioral Science*, 18 (3), 173–88.

Griffin, Ricky and K. M. Kacmar (1991), "Laboratory Research in Management: Misconceptions and Missed Opportunities," *Journal of Organizational Behavior*, 12 (4), 301–11.

Griffith, David E. and R. T. Rust (1994), "The Price of Competitiveness in Competitive Pricing," working paper available from Roland T. Rust, Owen Graduate School of Management, Vanderbilt University.

Hendon, Donald W. (1986), *Battling for Profits: How to Win Big on the Marketing Battlefield*. Jonesboro, AR: Business Consultants International.

Jacobson, Robert and D. A. Aaker (1985), "Is Market Share All That It's Cracked Up to Be?" *Journal of Marketing*, 49 (Fall), 11–21.

Kaplan, A. D. H., J. B. Dirlam, and R. F. Lanzillotti (1958), *Pricing in Big Business*. Washington, DC: The Brookings Institution.

Kiechel, Walter (1981), "The Decline of the Experience Curve," *Fortune*, (October 5), 139–46.

Kohn, Alfie (1986), *No Contest: The Case Against Competition*. Boston: Houghton Mifflin.

Kotler, Philip (1988), *Marketing Management: Analysis, Planning, Implementation and Control*, 6th ed. Englewood Cliffs, NJ: Prentice-Hall.

Kuhlman, D. Michael and A. F. J. Marshello (1975), "Individual Differences in Game Motivation as Moderators of Preprogrammed Strategy Effects in Prisoner's Dilemma," *Journal of Personality and Social Psychology*, 32 (5), 922–31.

Lanzillotti, Robert F. (1958), "Pricing Objectives in Large Companies," *American Economic Review*, 48 (5), 921–40.

Leeflang, Peter S. H. and D. R. Wittink (1996), "Competitive Reaction Versus Competitive Response: Do Managers Overreact?" *International Journal of Research in Marketing*" (forthcoming).

Lieberman, Marvin B. (1987), "The Learning Curve, Diffusion, and Competitive Strategy," *Strategic Management Journal*, 8, 441–52.

Liebrand, W. B. G. and G. J. van Run (1985), "The Effects of Social Motives on Behavior in Social Dilemmas in Two Cultures," *Journal of Experimental Social Psychology*, 21 (1), 86–102.

Locke, Edwin A. (1986), *Generalizing from Laboratory to Field Settings*. Lexington, MA: Lexington Books.

——— and G. P. Latham (1990), *A Theory of Goal Setting and Task Performance*. Englewood Cliffs, NJ: Prentice-Hall.

Messick, David M. and C. G. McClintock (1968), "Motivational Bases of Choice in Experimental Games," *Journal of Experimental Social Psychology*, 4 (1), 1–25.

——— and W. B. Thorngate (1967), "Relative Gain Maximization in Experimental Games," *Journal of Experimental Social Psychology*, 3 (1), 85–105.

Montgomery, Cynthia A. and B. Wernerfelt (1991), "Sources of Superior Performance: Market Share Versus Industry Effects in the U.S. Brewing Industry," *Management Science*, 37 (8), 954–59.

Mueller, Dennis C. (1992), "The Corporation and the Economist," *International Journal of Industrial Organization*, 10 (1), 147–69.

O'Leary-Kelly, Anne M., J. J. Martocchio, and D. W. Frink (1994), "A Review of the Influence of Group Goals on Group Performance," *Academy of Management Journal*, 37 (5), 1285–301.

Parks, Bill, S. W. Pharr, and B. D. Lockeman (1994), "A Marketer's Guide to Clausewitz: Lessons for Winning Market Share," *Business Horizons*, 37 (July-August), 68–73.

Porter, Michael E. (1980), *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. New York: The Free Press.

Prescott, John E., A. K. Kohli, and N. Venkatraman (1986), "The Market Share–Profitability Relationship: An Empirical Assessment of Major Assertions and Contradictions," *Strategic Management Journal*, 7 (4), 377–94.

Scodel, Alvin, J. S. Minas, P. Ratoosh, and M. Lipetz (1959), "Some Descriptive Aspects of Two-Person Non-Zero-Sum Games," *Journal of Conflict Resolution*, 3 (2), 114–19.

Shimp, Terence A., E. M. Hyatt, and D. J. Snyder (1991), "A Critical Appraisal of Demand Artifacts in Consumer Research," *Journal of Consumer Research*, 18 (3), 273–83.

Siegel, Sidney and N. J. Castellan, Jr. (1988), *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw Hill.

Sigall, H., E. Aronson, and T. van House (1970), "The Cooperative Subject: Myth or Reality?" *Journal of Experimental Social Psychology*, 6 (1), 1–10.

Slater, Robert (1993), *The New GE: How Jack Welch Revived an American Institution*. Homewood, IL: Business One Irwin.

Slater, Stanley F. and J. C. Narver (1994), "Does Competitive Environment Moderate the Market Orientation-Performance Relationship?" *Journal of Marketing*, 58 (January), 46–55.

——— and T. J. Zwirlein (1992), "Shareholder Value and Investment Strategy Using the General Portfolio Model," *Journal of Management*, 18 (4), 717–32.

Suls, Jerry M. (1977), "Social Comparison Theory and Research: An Overview from 1954," in *Social Comparison Processes*, Jerry M. Suls and R. L. Miller, eds. New York: John Wiley & Sons.

Szymanski, David M., S. G. Bharadwaj, and P. R. Varadarajan (1993), "An Analysis of the Market Share–Profitability Relationship," *Journal of Marketing*, 57 (July), 1–18.

Teger, Allan I. (1980), *Too Much Invested to Quit*. New York: Pergamon.

Tschoegl, Adrian E. and C. J. Yu (1990), "Gibrat's Law in Marketing: The Case of Liquor Brand Sales," *International Journal of Research in Marketing*, 7 (4), 249–62.

BUSINESS BEHAVIOR, VALUE AND GROWTH

Revised Edition

WILLIAM J. BAUMOL

PRINCETON UNIVERSITY

Harcourt, Brace & World, Inc.
New York / Chicago / San Francisco / Atlanta

13

WHARTON REPROGRAPHICS
81
259

2-4-6

2

ANTECEDENTS OF THE BEHAVIORAL THEORY OF THE FIRM[1]

In a modern market society, economic decisions on price, output, product lines, product mix, resource allocation, and other standard economic variables are made not by individual entrepreneurs but by a complex of private and public institutions. Many of these decisions are made within the large, multifunctional, and complicated organizations called firms. These are simple facts. They may not be facts with which economic theory should concern itself, but the disparity between the process by which business decisions appear to be made by complex organizations in the real world and the way in which they are explained by economic theory has provided material for several decades of debate.[2] In this chapter, we propose

to discuss the theory of the firm, the debates over it, and the relevance of existing theories of organizations to a revised theory of the firm. With such a discussion, we hope to articulate both our view of the limitations of existing theories and our motives for developing the behavioral theory of the firm outlined in this volume.

2.1 The Theory of the Firm

Any brief presentation of a body of theory is likely to suffer from the distortion dictated by condensation. This problem is severe when the theory has been a focus for controversy and the writers making the condensation are parties to the controversy. We are in such a position in describing the theory of the firm. Nevertheless, without completely concealing our prejudices, we describe in this section the current consensus on the theory of the firm, the challenges to that theory, and the rejoinders of the defenders of orthodoxy.

2.1.1 Current consensus on the theory of the firm

It will become clear below that there is disagreement about the theory of the firm in three respects. There is disagreement about (1) what the theory is, (2) the extent to which the theory is defective, and (3) appropriate methods for improving the theory.

The first of these disagreements makes it difficult to describe any current consensus on the theory of the firm. We are sympathetic with the view that, in fact, there is no consensus. At the same time, a case can be made for moderate agreement. We believe a fair number of economists would agree, except for questions of detail, on the general characterization of the theory of the firm given in this section.[3]

Assuming that the firm is operating within a perfectly competitive market, the generally received theory asserts that the objective of the firm is to maximize net revenue in the face of given prices and a technologically determined production function. Net revenue (profits, or expected profits) is the difference between receipts and the sum of fixed and variable costs. The production function is a relation between factors of production and their corresponding outputs determined by physical conditions within the firm. Maximization of profit is accomplished by determining the optimal mix of outputs (products) and inputs (factors), i.e., the equilibrium position.

---

[1] This chapter is based in part on a paper given in December, 1959, at the annual meeting of the Econometric Society and a paper published in *Contributions to Scientific Research in Management* (UCLA).

[2] There are a large number of articles that might be cited in this connection. Perhaps the articles stimulated by the so-called "marginalist" controversy are representative. See R. A. Lester, "Shortcomings of Marginal Analysis for Wage-Employment Problems," *American Economic Review*, 36 (1946), 63–82; "Marginalism and Labor Markets," *American Economic Review*, 37 (1947), 135–148; F. Machlup, "Marginal Analysis and Empirical Research," *American Economic Review*, 36 (1946), 519–554; "Rejoiner to an Antimarginalist," *American Economic Review*, 37 (1947), 148–154; R. A. Gordon, "Short-Period Price Determination," *American Economic Review*, 38 (1948), 265–288.

[3] See R. G. D. Allen, *Mathematical Economics* (London: St. Martins, 1957), 608–617; R. Dorfman, *Application of Linear Programming to the Theory of the Firm* (Berkeley: U. of Cal., 1951), 5–10; J. M. Henderson and R. E. Quandt *Microeconomic Theory* (New York: McGraw-Hill, 1958), 42–75; P. Samuelson, *Fou[illegible]ns of Economic Analysis* (Cambridge: Harvard University Press, 1947), 57–89.

Existing theory of the firm treats two main areas — the conditions for maximum net revenue, and the analysis of shifts in equilibrium positions. The usual method for obtaining a solution is first to derive the conditions for minimum cost at any fixed output and subsequently to determine the optimal level of output. This analysis, however, is usually limited to the single-product firm. A somewhat more general treatment is gained from the Hicks-Allen analysis of the multiproduct firm.

Using the methods of differential calculus, the Lagrangean multiplier, and the theory of quadratic forms, the propositions listed below are generally derived. The first two are given by a direct consideration of the sufficient conditions for an equilibrium.

1. n equations may be derived and solved for the optimal quantities of the firm's n commodities (both inputs and outputs).

2. At equilibrium, the marginal rate of substitution between two products, or between two factors, is equal to the ratio of their prices.

If we turn to variations near the equilibrium, the necessary conditions yield further relations.

3. The marginal physical productivity of a factor with respect to a product (the rate of change of the amount of the factor used with respect to the product's output) is equal to their price ratio.

4. The quantity of a good produced is selected so that its marginal cost (with respect to this product) is equal to its (given) price.

The sufficient conditions for equilibrium insure a maximum (rather than a minimum) of net revenue. From them the remainder of the theory is derived.

5. A price increase for a product raises its supply; a price increase for a factor reduces its demand.

6. "Cross" price effects are symmetric. That is, the rate of change of a first commodity with respect to the price of a second commodity is equal to the rate of change of the second with respect to the price of the first.

7. A price increase of a good tends to affect the other commodities by decreasing the outputs of products and increasing the inputs of factors.

Thus product-factor substitution is seen to dominate generally. However, within this framework complementarity with respect to a group of products *or* factors (a single price increase raising all their supplies *or* reducing all their demands, respectively) may operate.

In our judgment, the above description (see the appendix to this chapter for a mathematical description) is the closest thing to a consensus on the content of the theory of the firm. The theory has been extended to cover either imperfect factor markets, imperfect product markets, or both.[4] Although agreement on such extensions is somewhat less complete than for the perfectly competitive case, we can still describe them as "standard." These elaborations, while extending the theory to new market situations, retain the basic framework and decision-making process postulated for the firm in perfect competition.

Both the theory of monopolistic competition and attempts at oligopoly theory are such elaborations. In the theory of monopolistic competition, it is assumed that the firm no longer has a demand curve of infinite elasticity because its product is not a perfect substitute for the products of other firms in the industry. For example, Chamberlin's analysis is worked out essentially with three curves — a demand curve for the firm assuming the quality of the products of competitors and the prices of competitors' products given, a demand curve for the firm assuming that competitors' prices are always identical with the firm's price, and an average cost curve for the firm.[5] Predictions are derived from the model by assuming profit maximization by the firm and by assuming that each firm expects competitors not to follow its behavior (because competitors are so numerous, none is significantly affected by the behavior of a single firm). Within this framework Chamberlin handles three decision variables — price, product quality, and selling expense. From the standpoint of the theory of the firm, the theory of monopolistic competition adds two dimensions (selling expense and product quality) to the kinds of decisions the firm makes but leaves the decision process essentially unchanged from the traditional model.[6]

Similarly, efforts to develop a theory of oligopoly have generally retained the classical assumptions about firm behavior.[7] The theories focus on the interdependence among firms in oligopolistic markets, and this focus is characterized by an emphasis on such concepts as conjectural variations. A major difficulty in developing the theory is the assignment of proper values to conjectural variation terms.[8] Once meaning is given to the conjectural variation terms, the postulated decision process remains essentially the same as in the conventional model.[9]

---

[4] In recent years linear programing has been applied to the analysis of the firm. Linear programing can be viewed as an alternative optimizing technique to that of the differential calculus employed in the marginal analysis of the firm. From the standpoint of the present discussion, this innovation does not alter the basic nature of the theory.

[5] E. H. Chamberlin, *The Theory of Monopolistic Competition*, 5th ed. (Cambridge: Harvard University Press, 1946), 90.

[6] For another point of view see Chamberlin's "Full Cost and Monopolistic Competition," *Economic Journal*, 62 (1952), 318–325.

[7] G. J. Stigler, *The Theory of Price* (New York: Macmillan, 1946), 260.

[8] A. L. Bowley, *Mathematical Groundwork of Economics* (Oxford: Clarendon Press, 1924).

Both the theory of monopolistic competition and oligopoly theory have taken the basic theory of the firm as given and changed the market assumptions.[10] Dissatisfaction with this limited revision has led to a re-examination of the basic theory of the firm itself. In the debate over the theory of the firm, we can identify two major difficulties perceived by economists who view the basic theory as deficient. First, the motivational and cognitive assumptions of the theory appear unrealistic. Profit maximization, it is commonly alleged, is either only one among many goals of business firms or not a goal at all. As we shall see, this general criticism and its rejoinders have periodically filled the pages of economic journals with a host of subtleties both economic and philosophic. On the cognitive side, both the classical assumption of certainty and its modern equivalent — knowledge of the probability distribution of future events — have been challenged.

Second, the "firm" of the theory of the firm has few of the characteristics we have come to identify with actual business firms. It has no complex organization, no problems of control, no standard operating procedures, no budget, no controller, no aspiring "middle management." To some economists it has seemed implausible that a theory of an organization can ignore the fact that it is one.

In the next two sections, we discuss each of these criticisms and some of the forms each has taken, especially some of the proposals for revision. The objective is not to provide an exhaustive list of the large number of variations on the two themes but to suggest the nature of the criticisms and the problems the criticisms create in turn, for it is striking that the revisions have failed to secure widespread acceptance.

2.1.2 Challenges to the motivational and cognitive assumptions of the theory

The assumptions of rationality in the theory of the firm can be reduced to two propositions: (1) firms seek to maximize profits; (2) firms operate with perfect knowledge. Of course, a number of attempts have been made to adapt these assumptions to make them somewhat more sophisticated.

---

[9] It is as a solution to such problems that game theory seems plausible. However, the strategic decision rules of game theory thus far have offered more in formulation than they have been able to deliver in detailed economic models. Even the recent efforts in this direction by Shubik, although they develop economic game theory substantially, still leave a significant gap between the phenomena with which the theory would like to deal and the present power of *n*-person game theory. M. Shubik, *Strategy and Market Structure* (New York: Wiley, 1959). It is true, however, that the decision process is not identical with the conventional one when game theory is used. See A. Henderson, "The Theory of Di ly," *Quarterly Journal of Economics*, 68 (1954), 565–584.

[10] J. Robi *Economics of Imperfect Competition* (London: Macmillan, 1933).

Thus, we can assume that firms maximize the discounted value of future profits, and that firms have perfect knowledge only up to a probability distribution of all possible future states of the world. We are not concerned at this point with such adaptations, for we wish first to consider a set of more frontal assaults on the assumptions.

Proposed substitutes for profit maximization. There have been two basic challenges to the profit maximization assumption. First, is profit the only objective of business firms? Second, does maximization describe what the business firms do about the profits? As we shall see, some current critics link these two queries, but they need not be linked.

If not profit, what? Perhaps the simplest attack on profits as a motive is also the most destructive. We can argue that entrepreneurs, like anyone else, have a host of personal motives.[11] Profit is one, perhaps, but they are also interested in sex, food, and saving souls. It is rather difficult to deny the proposition, but if we accept it as critical, it is not easy to see how to devise a theory of the firm in anything approximating its present form (or even with its present goals). We can adopt the strategy of introducing subjective utility, but this cannot help us until the utile becomes a better-behaved measure. In a related fashion, we can argue (as Papandreou does) that organizational objectives grow out of interaction among the various participants in the organization.[12] This interaction produces a "general preference function." Whether we agree with Papandreou that this leaves us in no more vague a position than profit maximization, we must concede his point that little can be done with a general preference function unless it can be measured in some nontautological fashion (see Chapter 9).

An alternative to a utility-like preference function is substituting a different *summum bonum* for profits. For example, Rothschild has suggested that the primary motive of the entrepreneur is long-run survival.[13] In this view, decisions aim to maximize the security level of the organization (i.e., the probability that the organization will survive over the indefinite future). Another such suggestion is Baumol's, that firms seek to maximize sales subject to a profit constraint.[14]

The second attack on the assumption of profit maximization does not deny the importance of profits but questions the assumption of maximization. Gordon, Simon, and Margolis have all argued that profit maximizing should be replaced with a goal of making satisfactory profits.[15] Satisfactory profits represent a level of aspiration that the firm uses to evaluate alternative policies. The aspiration level may change over time, but in the short run it defines a utility function with essentially only two values — good enough and not good enough.

---

[11] G. Katona, *Psychological Analysis of Economic Behavior* (New York: McGraw-Hill, 1951).

[12] A. Papandreou, "Some Basic Problems in the Theory of the Firm," in *A Survey of Contemporary Economics*, ed. B. F. Haley (Homewood, Ill.: Richard D. Irwin, 1952), vol. 2, 183–219.

[13] K. W. Rothschild, "Price Theory and Oligopoly," *Economic Journal*, 42 (1947), 297–320.

[14] W. J. Baumol, *Business Behavior, Value and Growth* (New York: Mac , 1959), 45–53.

This critique of maximizing is linked with other proposals for revision in at least two ways. First, representing goals by a two-valued utility function avoids many of the problems involved in reducing the various dimensions of a multitude of goals to a single scale. Each goal enters as a simple constraint. All of the goals taken together define a space of acceptable solutions. Second, combining the acceptable level goal with a search theory (as found in Simon) suggests some possible solutions not only to the difficulties about motives, but also to objections about the cognitive side of the rationality assumptions.

Proposed substitutes for perfect knowledge. As we have already noted, it is awkward to assume perfect knowledge in the theory of the firm, and introducing expected value calculations in the case of risk solves only some of the problems. In particular, Simon and others have argued that information is not given to the firm but must be obtained, that alternatives are searched for and discovered sequentially, and that the order in which the environment is searched determines to a substantial extent the decisions that will be made.[16] In this way, the theory of choice and the theory of search become closely interwined and take on prime importance in a general theory of decision making.

A more conventional attempt to deal with the same problem is reflected in recent contributions to a normative theory of search.[17] In this theory, search activity is one of the competitors for internal resources, and expenditures for search are made up to the point where the marginal cost of search equals the marginal expected return from it. These elaborations of the theory are clearly more traditional than Simon's suggestions, but they do not seem to meet the burden of the complaint.

2.1.3 Challenges to the conventional concept of the firm

If we take seriously the concept of a firm as something distinct from an individual entrepreneur, there is no consensus on a theory of the firm.

---

[15] Gordon, *op. cit.;* J. Margolis, "The Analysis of the Firm: Rationalism, Conventionalism, and Behaviorism," *Journal of Business*, 31 (1958), 187–199; H. A. Simon, "A Behavioral Model of Rational Choice," *Quarterly Journal of Economics*, 69 (1952), 99–118.

[16] J. G. March and H. A. Simon, *Organizations* (New York: Wiley, 1958).

[17] A. Charnes and W. W. Cooper, "The Theory of Search: Optimum Distribution of Search Effort," *Management Science*, 5 (1958), 450–458.

As we have seen, the theory in all its standard forms ignores the fact that decisions are made in large-scale organizations. There is only one conspicuous exception to this statement. Marshall, as others before him, was impressed by the apparent increasing returns to scale.[18] He undertook to explain the historical reduction in production costs by introducing the concepts of internal and external economies, and it is this germ of an idea about the effect of organizational size on organizational performance, by the way of "internal economies," that became important to several decades of theory development.[19] This development, however, ignored other organizational aspects of the firm.

In recent years there has been speculation about the effects of other organizational factors. We will not try to describe all the suggestions for revision that have come from observing, or reflecting on, the behavior of actual business organizations. However, a sample of them will suggest the nature of the difficulties perceived in the theory and the kinds of revisions suggested.

First, the failure to view the firm as an organization has been criticized.[20] Papandreou has made perhaps the most detailed argument for expanding the framework of the firm.[21] He views the firm as a cooperative system. The executive tasks are accomplished by a "peak coordinator." The firm has certain goals, and it is the peak coordinator's job to achieve these by allocating resources rationally. This involves three actions: (1) substantive planning — constructing the firm's budget; (2) procedural planning — constructing a system of communication and authority; (3) executing both plans.

Within such a model of the firm, Papandreou sees certain areas of psychology playing a helpful role. The goals of the firm are strongly influenced by both internal and external forces. The internal influences come from such entities as stockholders, unions, government, and so forth. The preference function discussed above is a "resultant of the influences which are exerted upon the firm. The peak coordinator's job is to maximize the preference function." The greatest difficulty with the analysis is that it does not relate the model specifically to decision making. Hence, it provides a general analysis of the firm from the standpoint of organization theory without specifying precisely how the model can be used in economics.

---

[18] A. Marshall, *Principles of Economics*, 8th ed. (London: Macmillan, 1930), 278–290. See also G. J. Stigler, *Production and Distribution Theories* (New York: Macmillan, 1941), 76–83.

[19] P. Sraffa, "The Laws of Return Under Competitive Conditions," *Economic Journal*, 36 (1926), 535–550.

[20] W. W. Cooper, "Revisions to the Theory of the Firm," *American Economic Review*, 39 (1949), 1204–1222; G. F. Thirlby, "Notes on the Maximization Process in Company Administration," *Economica*, 17 (1950), 266–282.

[21] A. Papandreou, *op. cit.*

Second, the way in which the theory portrays the process of decision making in business firms has been criticized. It is alleged that firms, in fact, do not equate marginal cost and marginal revenue in deciding on either output or price; rather they follow one or another of a series of rules of thumb.

Consider on the price side, for example, the full cost (or mark-up) method of pricing.[22] Under this method, average variable cost is taken as a base and this cost is marked up to obtain the selling price. The mark-up presumably covers overhead costs and some profit. There are problems with this decision rule. There are unsettled questions as to how the size of the mark-up and the quantity of output are determined. However, the skepticism that greeted the original study of Hall and Hitch in 1939 seems to have been unwarranted. We now have evidence suggesting strongly that mark-up pricing is common in both England and the United States.

Consider similarly, on the output side, Eiteman's suggestion for using inventory turnover as a surrogate for demand estimation.[23] Inventory turnover then determines output. (As we have noted, the indeterminacy of output under mark-up pricing is bothersome.) Output is increased at the given price until inventories increase at abnormal rates. Output is then adjusted until inventories are stable.. Price is determined on a mark-up basis for the costs at that output which represents "the minimum reasonable use of the plant available." No doubt, in actual practice the role of inventory change in determining output is more complicated than this. However, if one drops the assumption that the firm knows with certainty its demand curve, or even a few points on it, it is clear that the firm must use feedback from the environment in fixing its price and output. Inventory change is one way of obtaining such feedback (normative inventory models largely build on this). Although there are other feedback devices, many firms do use information on inventory turnover in making the output decision.

Consider similarly the apparent sharp difference between the economic concept of costs on the one hand and the accounting concept of costs on the other. Cooper has outlined tentatively how standard costs and common control systems in business firms might be introduced into a meaningful theory of the firm.[24] It seems quite unlikely that firms use very frequently anything approaching the conventional marginal calculation of costs.

Consider finally some broader observations on the process of organizational decision making. As we observed earlier, Gordon[25] suggested the substitution of "satisfactory profits" for "profit maximization" as the goal of the firm. Even more important was his suggestion that standard business procedures influence the content of the decisions. For example, virtually every study of business decision making reveals the use of what Gordon called "shortcuts." These shortcuts, or guides to policy, take the form of dicta: "Thou shalt not allow thy share of the market to go below 10 per cent." "Honor thy dealers above all." "Thou shalt not unionize." Since the relation between these dicta and profit maximization is vague and the imperative "Optimize!" is less operational than the shortcuts, it is reasonable to expect that the so-called shortcuts will in fact dominate the decisions, at least in the short run.

Gordon extended his general process-oriented comments further. He argued that (1) the theory should reflect the way in which costs are actually used in a firm (e.g., selling expense); (2) the treatment of uncertainty in the theory is considerably at variance with the way in which firms react to uncertainty; (3) there is a critical discrepancy between theoretical time and real time; (4) the executive in the firm deals with only a subset of the decision variables specified in the theory — treating the others as essentially outside his control.

The details of Gordon's argument (and of similar more recent proposals) indicate that substantial differences have been found between the decision-making process of firms and the decision-making process of the theory of the firm.[26] They reflect a rather understandable discontent with what at least appears to be a very substantial disparity between theory and reality.

2.1.4 In defense of orthodoxy

The suggestions for revision have not gone unchallenged by the defenders of the classical theory. In fact, one writer or another has questioned the methodological relevance, the factual premises, or the theoretical utility of virtually all of the proposed revisions.

The methodological argument asserts simply that it is irrelevant whether the assumptions are valid. The only crucial test of a theory is its predictive power. The foremost proponent of this view is Friedman.[27] Friedman hits sharply at the view that the realism of assumptions is important. He maintains that concern over descriptive accuracy has led to a misunderstanding of economic theory. The function of economic theory is to build propositions with which the world can be analyzed and not to reproduce the world. The most crucial part of the Friedman argument, however, is also the most difficult to demonstrate. He argues that the current theory meets the predictive test, though the "evidence is extremely hard to document; it is scattered in numerous memorandums, articles and monographs concerned primarily with specific concrete problems rather than with submitting the hypothesis to test." (See Appendix A.)

---

[22] R. L. Hall and C. J. Hitch, "Price Theory and Business Behavior," in *Oxford Studies in the Price Mechanism*, ed. T. Wilson and P. W. S. Andrews (Oxford: Oxford Jniversity Press, 1951), 107–138.

[23] W. J. Eiteman, *Price Determination*, Bureau of Business Research Report No. 16 (Ann Arbor, Mich., 1949).

[24] W. W. [illegible]r, "A Proposal for Extending the Theory of the Firm," *Quarterly Journal of E*[illegible]*ics*, 65 (1951), 87–109.

[25] R. A. Gordon, *op. cit.*

[26] R. M. Cyert, H. A. Simon, and D. B. Trow, "Observation of a Business Decision," *The Journal of Business*, 29 (1956), 237–248.

[27] M. Friedman, "The Methodology of Positive Economics," in *Essays* [illegible]*ositive Economics* (Chicago: University of Chicago [illegible]

A second defense, not altogether consistent with the first, attempts to demonstrate the reality of some of the controversial assumptions underlying conventional theory. This approach has followed two lines. In one the theory has been interpreted to show that the assumptions are reasonable. Machlup's well-known paper represents the most elaborate defense of this type.[28] The important step he takes is to argue that cost and revenue are subjective.

> "It should hardly be necessary to mention that all the relevant magnitudes involved — cost, revenue, profit — are subjective — that is, perceived or fancied by the men whose decisions or actions are to be explained (the business men) — rather than 'objective' . . . . Marginal analysis of the firm should not be understood to imply anything but subjective estimates, guesses and hunches."

[illegible]
2011
hunch!s

Although this approach runs the danger of turning the theory into a tautology, it is, of course, an answer to those who criticize the cognitive assumptions of the theory.

The second line attempts to demonstrate empirically that the assumptions are valid. For example, Earley argues that modern accounting techniques such as direct costing (as well as the techniques of operations research) enable businessmen to behave as postulated in the theory of the firm. By use of an elaborate questionnaire he has gathered some data from which he infers that they generally do so.[29]

The third general method of defending conventional theory might be called "evolutionary analysis."[30] The argument concedes that firms do not (or at least may not) consciously pursue the marginal calculations specified in the theory, that firms may have many different goals, and that therefore there may not be a one-to-one relation between the decision process in the firm and the analytic tools used by economists to "describe" that process. However, since in the long run firms will survive only if they (by whatever

---

[28] F. Machlup, "Marginal Analysis and Empirical Research," *American Economic Review*, 36 (1946), 519–554.

[29] J. S. Earley, "Marginal Policies of 'Excellently Managed' Companies," *American Economic Review*, 46 (1956), 44–70.

[30] A. Alchian, "Uncertainty, Evolution and Economic Theory," *Journal of Political Economy*, 58 (1950), 211–221.

process) make the decisions dictated by economic theory, the theory will predict the behavior of viable firms. Since only the fittest survive, we need only a theory of the fit.

The fourth defense of conventional theory depends on a proposition in the technology of theory construction. It is argued that the proposed changes in the theory cannot be dealt with analytically.[31] Even though it might be desirable to substitute other assumptions for those currently found in economic theory, there is too large a gap between the proposals and the mathematical tools available for developing a theory. With the introduction of the high-speed electronic computer, this defense has been somewhat muted — or at least deflected to an examination of the problems of constructing and testing computer models. (See Appendix B.)

2.1.5 Some observations on the debate

The debates summarized in the previous sections seem to have remarkable powers of reincarnation.[32] They neither die nor fade away with much permanence. Instead, with each publication of new empirical evidence or a new theoretical treatise, the argument is resumed. We will leave the more esoteric features of this remarkable immortality to students of the sociology of knowledge and the metaphysics of wisdom. On a more mundane level, it seems to us that much of the controversy is based on a misunderstanding of the questions the conventional theory of the firm was designed to answer. The theory of the firm, which is primarily a theory of markets, purports to explain at a general level the way resources are allocated by a price system. To the extent to which the model does this successfully, its gross assumptions will be justified. However, there are a number of important and interesting questions relating specifically to firm behavior that the theory cannot answer and was never developed to answer, especially with regard to the internal allocation of resources and the process of setting prices and outputs.

Thus, many of the attacks on the theory of the firm are not so much proper critiques of existing theory as they are suggestions for the development of a theory appropriate to this different set of questions. Ultimately, a new theory of firm decision-making behavior might be used as a basis for a theory of markets, but at least in the short run we should distinguish between a theory of microbehavior, on the one hand, and the microassumptions appropriate to a theory of aggregate economic behavior on the other. In the present volume we will argue that we have developed the rudiments of a reasonable theory of firm decision making. Our arguments for the utility of the theory in developing a theory of markets are, at least at the moment, more modest.

---

[31] A. Papandreou and J. Wheeler, *Competition and Its Regulation* (Englewood Cliffs, N.J.: Prentice-Hall, 1954), 73–74.

[32] T. C. Koopmans, *Three Essays on the State of Economic Science* (New York: McGraw-Hill, 1957), 137–142; E. T. Penrose, "Biological Analogies in the Theory of the Firm," *American Economic Review*, 42 (1952), 804–810; E. Rotwein, "On 'The Methodology of Positive Economics'," *Quarterly Journal of Economics*, 73 (1959), 554–575.

Thus, for present purposes, three characteristics of the conventional theory of the firm need emphasis:

1. It deals with a particular set of decisions (e.g., price, production) that are viewed as functions of a few "catch-all" variables (e.g., demand, costs).
2. The theory uses aggregation as a tactic; it attempts to specify total market supply and demand curves.
3. There is no attention to, or interest in, the actual process by which individual firms reach decisions.

Despite the record of controversy over the theory, these attributes are more conspicuous than bad. However, they will (in conjunction with a similar characterization of organization theory below) suggest some reasons why little success has been realized in the efforts to merge economic and organization theories of the firm.

2.2 Organization Theory

Although there have been a number of recent efforts to bring some rough order to the field,[33] the study of organizations is even more diffuse than the study of the economic activities of the firm; as a result, it is harder to characterize briefly. To avoid some of the difficulties, we limit our attention to one aspect of organization theory — its possible relevance to a revised theory of the firm. In a sense, what many of the critics of the theory of the firm propose is an integration of organization theory and the theory of the firm. On occasion, such a proposal has been made explicit, and some efforts to achieve an integration have been made.[34] We can consider whether such a proposal is reasonable.

Like the theory of the firm, organization theory means different things to different people. Only those meanings of the term that emphasize the empirical study of behavior in organizations are relevant to our present purpose. If we thus limit the focus, there are (as we view it) three major

---

[33] J. G. March and H. A. Simon, *op. cit.*; M. Haire, ed., *Modern Organization Theory* (New York: Wiley, 1959); A. Etzioni, *Complex Organizations* (New York: Holt, Rinehart and Winston, 1961).

[34] H. Le[illegible]stein, *Economic Theory and Organizational Analysis* (New York: Harper, 1960).

branches of organization theory. The first of these, which can best be described as "sociological," lists as its founding fathers Weber, Durkheim, Pareto, and Michels and centers on phenomena of bureaucracy. The second is "social psychological" and has been built primarily on an experimental base with an emphasis on an "efficiency" criterion. The third is "administrative" in the sense that it focuses on the problems of the executive in dealing with an organization and is centered on classic administrative axioms.

The early sociological theories of organization (e.g., Weber, Durkheim) emphasized the phenomena of division of labor and specialization as broad social trends and the importance of large-scale organizations in utilizing specialized competences. Weber, for example, placed considerable emphasis on the rationality of bureaucratic organization. Such theorists tended to compare the depersonalized professionalization of bureaucracies with a (probably overdrawn) view of earlier personal and unspecialized systems of social organization. To a certain extent the early theorists and, to a much greater extent, modern sociological students of organizations emphasized what Merton has labeled the "unanticipated consequences of purposive social action." Thus, the major variables tend to be such things as subgoal differentiation and conflict, individual personality changes, and organizational life-cycle.[35]

Social psychological approaches to organizational phenomena have tended to be less grandiose in scope. In general, they have taken a relatively obvious criterion of efficiency (e.g., productivity, speed) in a relatively simple task and examined experimentally the effect of some small set of independent variables on the efficiency of the organization. In this tradition have been the studies of communication nets, simulated radar warning stations, and small problem-solving groups. Somewhat less experimental have been the studies of morale and productivity, which also emphasize explicitly a criterion of efficiency.[36]

The final branch of organization theory dates in a sense from the earliest political and social philosophers. Speculation about centralization and decentralization and the problems of coordination can be found in pre-Christian writings. In this case, however, modern administrative theorists generally reject the earlier formulations with considerable vigor. In particular, Barnard and Simon have argued against the excessively formalistic and nonoperational analyses of early administrative theory.

Much of the recent work in this branch of organization theory takes the decision-making process as its specific focus of concern. The theory tends to subdivide, then, into two main parts, each concerned with a certain class of decisions of significance to organizations. The first subdivision views the organization as a clearing house through which transfer payments are arranged among participants (e.g., workers, investors, customers). The theory describes the "decision to belong," and specifies the conditions of organizational survival in terms of the methods of motivating organizational participants.[17]

---

[35] P. Blau, *Dynamics of Bureaucracy* (Chicago: University of Chicago Press, 1955); A. W. Gouldner, *Patterns of Industrial Bureaucracy* (Glencoe, Ill.: Free Press, 1954).

[36] C. Argyris, *Understanding Organizational Behavior* (Homewood, Ill.: Dors[illegible]960); R. Likert, *New Patterns of Management* (New York: McGraw-Hill, 1961).

The second subdivision aims at the development of theory explaining how decisions are made in organizations, with special emphasis upon the processes of executive influence and the impact of organizational position on individual goals and perceptions.[18]

Without attempting to describe the various forms that theoretical arguments have assumed within organization theory, we think this brief sketch of the focus of attention is enough to illustrate three major points:

1. The theory focuses on a set of problems that are different from those of the economic theory of the firm. Its problems are not specifically economic; virtually nothing is said about how output levels are set, advertising expenditures determined, and so forth.
2. Although it places considerable emphasis on the study of "process" — the study of what goes on in an organization — only the third branch of the theory focuses primarily upon organizational decision-making processes.
3. Unlike the theory of the firm, there is no consideration of "aggregation." Indeed, there is nothing to aggregate.

Thus, existing organization theory provides only a very partial basis for a new theory of the firm. The sociological and social psychological approaches have emphasized questions that are only marginally relevant to either the objectives of conventional theories of the firm or the objective of predicting individual firm behavior. The decision-making approach has developed a substantial theory of decision-making processes in an organizational context, but has not applied the theory to the specific environmental conditions in which the business firm operates nor applied the theory in detail to the particular decision variables that characterize the firm's operation.

---

[17] H. A. Simon, *Administrative Behavior* (New York: Macmillan, 1947); C. I. Barnard, *Functions of the Executive* (Cambridge: Harvard University Press, 1938); H. A. Simon, D. W. Smithburg, and V. A. Thompson, *Public Administration* (New York: Knopf, 1950). This subdivision of organization theory has many points of contact with the social psychological — largely with the motivational aspects of decisions.

[18] C. I. Barnard, *op. cit.*; H. A. Simon, *op. cit.* This subdivision has derived its basic framework from cognitive psychology [H. A. Simon, *Models of Man* (New York: Wiley, 1957), introduction to pt. IV; J. G. March and H. A. Simon, *op. cit.*, chaps. 6 and 7] but historically has derived its problems primarily from economics — the theory of the firm (Barnard), and public expenditure theory [C. E. Ridley and H. A. Simon, *Measuring Municipal Activities* (Chicago: International City Managers' Association, 1938)].

2.3 Research Questions in a Revised Theory of Firm Decision Making

Given these characteristics of organization theory — or much of it — and the theory of the firm — or much of it, it is clear that much more is needed than a simple integration of the two. *Our conception of the task we face is that of constructing a theory that takes (1) the firm as its basic unit, (2) the prediction of firm behavior with respect to such decisions as price, output, and resource allocation as its objective, and (3) an explicit emphasis on the actual process of organizational decision making as its basic research commitment.* The remainder of this book is directed toward that task. The task involves two main problems: determining the major attributes of decision making by business firms and identifying a theoretical language appropriate to a revised theory. We defer a discussion of the language problem until we attempt to describe explicit models of organizational decision making (Chapters 5, 7, 8, and 10). Assuming that we can solve the language problem, the substantive problem becomes pressing. If we can afford the luxury of greater complexity in the theory, to what questions do we require answers?

Much of this book is aimed at answering this last question. At the outset, however, we can anticipate the direction of the research. We start with a simple conception that an organizational decision is the execution of a choice made in terms of objectives from among a set of alternatives on the basis of available information. This leads to an examination of how organizational objectives are formed, how strategies are evolved, and how decisions are reached within those strategies.

Organizational objectives of a business firm. Once we drop the concept of a single, universal, organizational goal (e.g., profit maximization) and look instead at the process for defining objectives in organizations, we need some propositions about the development of goals. What is the effect of departmental structure on the goals actually pursued in an organization? It is commonly alleged that one of the most frequent phenomena in an organization is the differentiation of subunit goals and the identification of individuals with the goals of the subunits, independently of the contribution of that goal to the organization as a whole. What difference does this make for a business firm and its decisions on such matters as resource allocation?

What effect do planning and plans have on organizational objectives? Plans are a conspicuous part of organizational functioning, yet their impact on organizational goals is little understood. We would especially emphasize the study of the role of the budget in determining goals. Thus, we would argue that more information is necessary about how budgets are determined.

How do objectives change? If we allow goals to develop within the organization, we must also allow them to change over time. To what extent can we deduce propositions about shifts in organizational objectives from the theories and research on individual aspiration levels?

What characterizes the objectives that are defined in an organization? Assuming that profit maximization is sometimes replaced by the use of profit surrogates, acceptable-level criteria, or other goals, can we predict under what conditions a given objective will be used? Even more awkward, how do we deal with (i.e., what are the consequences of) conflict of interest within organizations?

Decision strategies. We are assuming that the decisions of the firm are not always uniquely determined by its external environment — for example, the market. Where the firm has discretion, we think it forms what we have called *decision strategies* or *rules*. Many important aspects of those strategies turn on the distinction made by the firm between decision variables and parameters. What is taken as given and what is treated as subject to manipulation by the firm? Here, a number of organizational characteristics seem to us to be important. Are some things treated as variable only at specified points in time? What difference does such time spacing of decisions make? To what extent are one department's decision variables another department's parameters and what difference does it make?

Decision making within strategies. Once organizational objectives and decision strategies are determined, the organization can be viewed as an information-processing and decision-rendering system. We need more reliable information on where and how organizations secure information, how that information is communicated through the organization, how authoritative decisions are reached, and finally, how such decisions are implemented in the organization.

To what extent does the market serve as a source of information about consumers, competitors, and so forth? What alternative sources of data exist and are used? What determines which sources of information will be used by an organization? Specifically, what predictable informational biases can be identified? If it is true that typically only a very small portion of the total available information is ever recorded by the organization, the processes by which the initial screening takes place has extraordinary importance in determining the final decision.

What happens to information as it is processed through the organization? What predictable screening biases are there in an organization? What is the effect of conflict of interest on communication? What difference does time pressure make?

What characterizes authoritative decision makers under different conditions [illegible]w do their prior experience and reference group identifications affect their decisions? How are executive expectations determined? What differences are there between individual and group decision making? How do hierarchical groups make decisions? If gradual commitment distinguishes decision making in business firms, in what respect does this affect the content of the decisions?

What is the relation between decisions made by the responsible executives and the final "decision" implemented by the organization? How do variations in the control systems (e.g., accounting systems) affect the implementation of decisions? What is the function and consequence of "organizational slack"? In what systematic ways are decisions elaborated or changed by the organization?

In general, there are many questions about the behavior of business firms but only a few answers. Existing theory is not equipped to answer most of the questions we have raised. Where an answer can be derived by brute force, it tends to be ambiguous or conspicuously inadequate. In order to develop an alternative theory, we need more satisfactory theories of organizational *goals*, organizational *expectations*, organizational *choice*, and organizational *control*. In our view, these are the four major subtheories of a behavioral theory of the firm. A theory of organizational goals would consider how goals arise in an organization, how they change over time, and how the organization attends to them. A theory of organizational expectations would treat how and when an organization searches for information or new alternatives and how information is processed through the organization. A theory of organizational choice would characterize the process by which the alternatives available to the organization are ordered and a selection made among them. A theory of organizational control would specify the differences between executive choice in an organization and the decisions actually implemented.

2.4 Summary

We have suggested that there are serious problems in using the theory of the firm and organization theory to predict and explain the behavior of business firms with respect to such economic decisions as price, output, capital investment, and internal resource allocation. In the remaining chapters we attack these problems in three ways:

1. We indicate partial answers to some of the questions posed in the previous section. Thus, in Chapter 3 we consider organizational goals; in Chapter 4, organizational expectations; in Chapter 5, organizational choice and (to a limited extent) organizational control.

2. We present some explicit models of organizational decision m[illegible]g. In Chapters 7 and 10 models of specific organizations are describ[illegible] In

Chapters 8 and 9 we indicate two general models of price and output determination.

3. In Chapter 11 we suggest some implications of the revised theory for future work, on both the micro and the macro levels.

Appendix. Mathematical Theory of the Firm[19]

The derivation of the seven propositions of Chapter 2 is shown in this appendix.

1. The production function for a multiproduct firm is given by

$$f(x_1, x_2, \ldots, x_n) = 0$$

where x_r $(r = 1, \ldots, n)$ is the output of the rth commodity. If $x_r > 0$, the commodity is a product, and if $x_r < 0$, it is a factor. Assume p products and $n - p$ factors.

Within a perfectly competitive market, the commodity prices p_r $(r = 1, \ldots, n)$ are given to the firm. It must then choose the optimal mix to maximize net revenue:

$$R = \sum_r p_r x_r - A$$

subject to

$$f(x_1, \ldots, x_n) = 0$$

where A is fixed cost.

The equilibrium position is determined using the Lagrangean multiplier λ by maximizing the new function

$$Z = R - \lambda f$$

for the n variables, x_r $(r = 1, \ldots, n)$.

The necessary conditions give

$$\frac{\partial Z}{\partial x_r} = p_r - \lambda f_r = 0$$

where $f_r = \partial f / \partial x_r$, i.e.,

$$\left.\begin{aligned} p_r &= \lambda f_r \\ \text{and} \qquad f(x_1, \ldots, x_n) &= 0 \end{aligned}\right\} \qquad (r = 1, \ldots, n) \tag{1}$$

as $(n + 1)$ equations to determine the quantities of the n commodities and λ.

---

[19] We are indebted to M. A. H. Dempster for this way of summarizing the literature.

2. From Eq. (1), the marginal rate of substitution between the rth and the sth product, or the rth and the sth factor, becomes

$$\frac{f_r}{f_s} = \frac{p_r}{p_s} \qquad \begin{matrix} (r, s = 1, \ldots, p) \\ \text{or} \\ (r, s = p + 1, \ldots, n) \end{matrix}$$

3. The marginal physical productivity, with respect to the rth product, of the sth factor,

$$M_{r,s} = \frac{\partial(-x_s)}{\partial x_r} = \frac{f_r}{f_s} = \frac{p_r}{p_s} \qquad \frac{(r = 1, \ldots, p)}{(s = p + 1, \ldots, n)}$$

where $\partial(-x_s)/\partial x_r$ is the rate of change of the amount of the factor used $(x_s < 0)$ with respect to the product.

4. Total costs are given by

$$C = \sum_s p_s(-x_s) + A \qquad (s = p + 1, \ldots, n)$$

such that the marginal cost of the rth good produced,

$$\frac{\partial C}{\partial x_r} = \sum_s p_s \frac{\partial(-x_s)}{\partial x_r} = \sum_s p_s M_{r,s} = \sum_s p_s \frac{p_r}{p_s} = p_r$$

5. The sufficient conditions for a maximum of net revenue, R, are

$$d^2Z = -\lambda\, d^2f < 0$$

This is equivalent to

$$d^2f = \sum_r \sum_s f_{rs}\, dx_r\, dx_s > 0$$

(i.e., d^2f is positive definite) subject to

$$df = \sum_r f_r\, dx_r$$

From the theory of quadratic forms the sufficient conditions may be written in the following manner:

$$\begin{vmatrix} 0 & f_1 & f_2 \\ f_1 & f_{11} & f_{12} \\ f_2 & f_{12} & f_{22} \end{vmatrix} < 0; \quad \ldots; \quad F = \begin{vmatrix} 0 & f_1 & f_2 & \ldots & f_n \\ f_1 & f_{11} & f_{12} & \ldots & f_{1n} \\ f_2 & f_{12} & f_{22} & \ldots & f_{2n} \\ \ldots & \ldots & \ldots & f_{rs} & \ldots \\ f_n & f_{1n} & f_{2n} & \ldots & f_{nn} \end{vmatrix} < 0$$

where

$$f_r = \frac{\partial f}{\partial x_r}, \qquad f_{rs} = \frac{\partial^2 F}{\partial x_r\, \partial x_s}$$

These can be interpreted as stability conditions for arbitrary finite variations x_r, $\dot{x}_s$. Note that F and all its principal minors are symmetric.

Alternatively,

$$\sum_r \sum_s \frac{F_{rs}}{F} y_r y_s > 0 \tag{2}$$

where the y's are any finite variations (not all zero) from the equilibrium position of the quantities of products or factors. F_{rs} is the cofactor of f_{rs} in F.

The effect of a variation in any one price, p_r, is determined by differentiating (1) to yield $(n+1)$ equations in $(n+1)$ variables,

$$\left(\frac{1}{\lambda}\frac{\partial\lambda}{\partial p_r}\right) \quad \text{and} \quad \frac{\partial x_s}{\partial p_r} \qquad (s = 1, \ldots, n)$$

Thus,

$$\sum_s f_s \frac{\partial x_s}{\partial p_r} = 0$$

$$f_i\left(\frac{1}{\lambda}\frac{\partial\lambda}{\partial p_r}\right) + \sum_s f_{is}\frac{\partial x_s}{\partial p_r} = 0 \qquad (i = 1, \ldots, r-1, r+1, \ldots, n)$$

$$f_r\left(\frac{1}{\lambda}\frac{\partial\lambda}{\partial p_r}\right) + \sum_s f_{rs}\frac{\partial x_s}{\partial p_r} = \frac{1}{\lambda}$$

This linear system is solved by Cramer's rule to yield

$$\frac{\partial x_s}{\partial p_r} = \frac{F_{rs}}{\lambda F} \qquad (r, s = 1, \ldots, n) \tag{3}$$

Substituting in (2) (since λ positive),

$$\sum_r \sum_s \frac{\partial x_s}{\partial p_r} y_r y_s > 0$$

For all y's zero except y_r,

$$\frac{\partial x_r}{\partial p_r} > 0 \qquad (r = 1, \ldots, n)$$

If $x_r > 0$, $\partial x_r/\partial p_r > 0$, and if $x_s < 0$, $\partial(-x_s)/\partial p_r < 0$.

6. Since F is symmetric, $F_{rs} = F_{sr}$ and (from (3))

$$\frac{\partial x_s}{\partial p_r} = \frac{\partial x_r}{\partial p_s}$$

7. The effects of a change in the price of the rth commodity, p_r, are determined by the expansion rule for determinants operating on F, and (1):

$$0 = \sum_s f_s F_{rs} = F \sum_s p_s\left(\frac{F_{rs}}{\lambda F}\right) = F \sum_s p_s \frac{\partial x_s}{\partial p_r}$$

Now, $F < 0$; hence

$$\sum_s p_s \frac{\partial x_s}{\partial p_r} = 0 \qquad (r, s = 1, \ldots, n)$$

and

$$\sum_s{}^{*} p_s \frac{\partial x_s}{\partial p_r} = p_r\left(-\frac{\partial x_r}{\partial p_r}\right) < 0 \tag{4}$$

where $\sum_s{}^{*}$ sums over all $s \neq r$.

Note that the negative values of the $(\partial x/\partial p_r)$'s on the left-hand side of (4) outweigh the positive.

3

ORGANIZATIONAL GOALS[1]

3.1 The Problem of Collective Goals

If we wish to develop a theory that predicts and explains business decision-making behavior, we face a problem that can be paraphrased in terms of the following:

1. People (i.e., individuals) have goals; collectivities of people do not.
2. To define a theory of organizational decision making, we seem to need something analogous — at the organization level — to individual goals at the individual level.

For the moment, let us accept this paraphrase (not everyone does). The theorist's problem is then to identify some concept of organization goals that is consistent with the apparent denial of their existence. Since (rightly or wrongly) individual goals are perceived as lodged in the individual human mind, the problem is to specify organizational goals without postulating an "organizational mind." In order to solve the problem we must (1) specify what an organization is in terms of the individual-organization dichotomy, (2) agree on the nature of the theoretical problems created by such a conception for the notion of organizational goals, and (3) identify a plausible solution to the theoretical problems. We will argue that there is substantial agreement on the first two requirements, but that the classic procedures for meeting the third requirement are deficient.

---

[1] This chapter is based extensively on R. M. Cyert and J. G. March, "A Behavioral Theory of Organizational Objectives," *Modern Organization Theory*, ed. M. Haire (New York: Wiley, 1959).

3.1.1 Conception of an organization

Let us view the organization as a coalition. It is a coalition of individuals, some of them organized into subcoalitions. In a business organization the coalition members include managers, workers, stockholders, suppliers, customers, lawyers, tax collectors, regulatory agencies, etc. In the governmental organization the members include administrators, workers, appointive officials, elective officials, legislators, judges, clientele, interest group leaders, etc. In the voluntary charitable organization there are paid functionaries, volunteers, donors, donees, etc.

Drawing the boundaries of an organizational coalition once and for all is impossible. Instead, we simplify the conception by focusing on the participants in a particular "region" — either temporal or functional. That is, over a specified (relatively brief) period of time we can identify the major coalition members; or, for a particular decision we can identify the major coalition members. More generally, for a certain class of decisions over a relatively long period of time we can specify the major classes of coalition members. As a result, we will be able to develop models of organizational decision making (for the short run) that pay only limited attention to the process by which the coalition is changed; but any such simplification involves some clear risks when we generalize to long-run dynamics.

This conception of an organization fits a number of recent formulations: the inducements-contributions schema, game theory, and the theory of teams. Each of these theories assumes a coalition of participants; each (with the exception of the inducements-contributions schema). assumes that by some procedure the coalition arrives at a statement of organization goals. However, the idea of an organization goal and the conception of an organization as a coalition are implicitly contradictory. Basic to the idea of a coalition is the expectation that the individual participants in the organization may have substantially different preference orderings (i.e., individual goals). That is to say, any theory of organizational goals must deal successfully with the obvious potential for internal goal conflict inherent in a coalition of diverse individuals and groups.

3.1.2 Classic devices for defining organization goals

There are two classic economic solutions to the problem of organization goals. The first, or entrepreneurial, solution is to describe an organization as consisting of an entrepreneur (either the top of the managerial hierarchy or some external control group such as stockholders) and a staff. The goals of the organization are then defined to be the goals of the entrepreneur. Conformity to these goals is purchased by payments (wages, interest, love) made by the entrepreneur to the staff and by a system of internal control that informs the staff of the entrepreneurial demands. This solution to the problem is characteristic of the economic theory of the firm, some political theories of public bureaucracies, and most theories of management.

The second solution to the problem is to identify a common or consensual goal. This is a goal that is shared by the various participants in the organization. It may be *a priori* sharing, as in many theories of political institutions in which the goal of "public interest" or "social welfare" is introduced. It may be *a posteriori* sharing, as in some theories of small-group goal formation through discussion. In either case, conflict is eliminated through consensus.

Neither solution is entirely happy. Both attempt to define a joint preference ordering for the coalition. In our view, such attempts are misdirected. Actual organizational goals cannot normally be described in terms of a joint preference ordering. Studies of organizational objectives suggest that agreement on objectives is usually agreement on highly ambiguous goals.[2] Such agreement is undoubtedly important to choice within the organization, but it is far from the clear preference ordering usually assumed. The studies suggest further that behind this agreement on rather vague objectives there is considerable disagreement and uncertainty about subgoals, that organizations appear to be pursuing different goals at the same time.[3] Finally, the studies suggest that most organization objectives take the form of an aspiration level rather than an imperative to "maximize" or "minimize," and that the aspiration level changes in response to experience.[4]

Unless we choose to ignore such observations, we need to reconsider our conceptions of objectives. Since the existence of unresolved conflict is a conspicuous feature of organizations, it is exceedingly difficult to construct a useful positive theory of organizational decision making if we insist on internal goal consistency. As a result, recent theories of organization objectives describe goals as the result of a continuous bargaining-learning process. Such a process will not necessarily produce consistent goals.

---

[2] D. B. Truman, *The Governmental Process* (New York: Knopf, 1951); A. D. H. Kaplan, J. B. Dirlam, and R. F. Lanzillotti, *Pricing in Big Business* (Washington, D.C.: Brookings Institution, 1958).

[3] Kaplan, *et al., op. cit.;* P. Selznick, *TVA and the Grass Roots* (Berkeley: University of California Press, 1949).

[4] P. M. Blau, *The Dynamics of Bureaucracy* (Chicago: University of Chicago Press, 1955); R. M "The Internal Organization of the Firm and Price Formation: An Illustrative *Quarterly Journal of Economics*, 63 (1949), 92–110.

3.2 The Goal Formation Process

In the theory to be outlined here, we consider three major ways in which the objectives of a coalition are determined. These are:

1. the bargaining process by which the composition and general terms of the coalition are fixed;
2. the internal organizational process of control by which objectives are stabilized and elaborated;
3. the process of adjustment to experience by which coalition agreements are altered in response to environmental changes.

3.2.1 Formation of coalition objectives through bargaining

A basic problem in developing a theory of coalition formation is that of handling side payments. It seems certain that the side payments by which organizational coalitions are formed do not satisfy the requirements of unrestricted transferability of utility. Side payments are made in many forms: money, personal treatment, authority, organization policy, and so forth. A winning coalition does not have a fixed booty that it then divides among its members. Quite to the contrary, the total value of side payments available for division among coalition members is a function of the composition of the coalition, and the total utility of the actual side payments depends on the distribution made within the coalition. There is no conservation of utility. For example, if we can imagine a situation in which any dyad is a viable coalition (e.g., a partnership to exploit the proposition that two can live more cheaply in coalition than separately), we would predict a greater total utility for those dyads in which needs were complementary than for those in which they were competitive.

Such a situation makes game theory as it currently exists virtually irrelevant for a treatment of organizational side payments,[5] but the problem is in part even deeper than that. The second requirement of theories such as game theory, theory of teams, and inducements-contributions theory is that after the side payments are made, a joint preference ordering is defined. All conflict is settled by the side-payment bargaining. The employment-contract form of these theories, for example, assumes that the entrepreneur has an objective. He then purchases whatever services he needs to achieve the objective. In return for such payments employees contract to perform whatever is required of them — at least within the range of permissible requirements. For a price, the employee adopts the "organization" goal.

One feature of such a conception is that it describes a coalition asymmetrically. To what extent is it arbitrary, in conventional accounting procedures, that we call wage payments "costs" and dividend payments "profit" rather than the other way around? Why is it that in our quasi-genetic moments we are inclined to say that in the beginning there was a manager and he recruited workers and capital? For the development of our own theory we make two major arguments. First, the emphasis on the asymmetry has seriously confused the understanding of organizational goals. The confusion arises because ultimately it makes only slightly more sense to say that the goal of a business organization is to maximize profit than to say that its goal is to maximize the salary of Sam Smith, Assistant to the Janitor.

---

[5] D. Luce and H. Raiffa, *Games and Decisions* (New York: Wiley, 1957), . 7 and 10.

Second, despite this, there are important reasons for viewing some coalition members as quite different from others. For example, it is clear that employees and management make somewhat different demands on the organization. In their bargaining, side payments appear traditionally to have performed the classical function of specifying a joint preference ordering. In addition, some coalition members (e.g., many stockholders) devote substantially less time to the particular coalition under consideration than do others. It is this characteristic that has been used to draw organizational boundaries between "external" and "internal" members of the coalition. Thus, there are important classes of coalition members who are passive most of the time. A condition of such passivity is that the payment demands they make be of such a character that most of the time they can be met rather easily.

Although we thereby reduce substantially the size and complexity of the coalition relevant for most goal setting, we are still left with something more complicated than an individual entrepreneur. It is primarily through bargaining within this active group that what we call *organizational objectives* arise. Side payments, far from being the incidental distribution of a fixed, transferable booty, represent the central process of goal specification. That is, a significant number of these payments are in the form of policy commitments.

The distinction between demands for monetary side payments and demands for policy commitments seems to underlie management-oriented treatments of organizations. It is clear that in many organizations this distinction has important ideological connotations. Indeed, the breakdown of the distinction in our generation has been quite consistently violent. Political party machines in this country have changed drastically the ratio of direct monetary side payments (e.g., patronage, charity) to policy commitments (e.g., economic legislation). Labor unions are conspicuously entering into what has been viewed traditionally as the management prerogatives of policy making and demanding payments in that area. Military officers have given up the substance — if not entirely the pretense — of being simply hired agents of the regime. The phenomenon is especially obvious in public[6] and voluntary[7] organizations, but all organizations use policy side payments. The marginal cost to other coalition members is typically quite small.

This trend toward policy side payments is especially noticeable in contemporary organizations, but the important point is that we have never come close to maintenance of a sharp distinction in the kinds of payments made and demanded. Policy commitments have (one is tempted to say *always*) been an important part of the method by which coalitions are formed. In fact, an organization that does not use such devices can exist in only a rather special environment.

To illustrate coalition formation under conditions where the problem is not scarce resources for side payments but varying complementarities of policy demands, imagine a nine-man committee appointed to commission a painting for the village hall. The nine members make individually the following demands:

A: The painting must be an abstract monotone.
B: The painting must be an impressionistic oil.
C: The painting must be small and oval in shape.
D: The painting must be small and in oil.
E: The painting must be square in shape and multicolored.
F: The painting must be an impressionistic square.
G: The painting must be a monotone and in oil.
H: The painting must be multicolored and impressionistic.
I: The painting must be small and oval.

In this case each potential coalition member makes two simple demands. Assuming that five members are all that are required to make the decision, there are three feasible coalitions: A, D, C, G, and I can form a coalition and commission a small, oval, monotone, oil abstract; B, C, D, H, and I can form a coalition and commission a small, oval, multicolored, impressionistic oil; B, D, E, F, and H can form a coalition and commission a small, square, multicolored, impressionistic oil.

Committeeman D, it will be noted, is in the admirable position of being included in every possible coalition. The reason is clear; his demands are completely consistent with the demands of everyone else.

Obviously, at some level of generality the distinction between money and policy payments disappears because any side payment can be viewed as a policy constraint. When we agree to pay someone $35,000 a year, we are constrained to the set of policy decisions that will allow such a payment. Any allocation of scarce resources (such as money) limits the alternatives for the organization. However, the scarcity of resources is not the only kind of problem. Some policy demands are strictly inconsistent with other demands. Others are completely complementary. To be sure, the problems of policy consistency are in principle amenable to explicit optimizing behavior, but they add to the computational difficulties facing the coalition members and make it even more obvious why the bargaining leading to side payment and policy agreements is only slightly related to the bargaining anticipated in a theory of omniscient rationality.

---

[6] R. A. Dahl and C. E. Lindblom, *Politics, Economics, and Welfare* (New York: Harper, 1953); H. A. Simon, D. W. Smithburg, and V. A. Thompson, *Public Administration* (New York: Knopf, 1950).

[7] D. L. Sills, *The Volunteers* (Glencoe, Ill.: Free Press, 1957); S. L. Messinger, "Organizational Transformation: A Case Study of a Declining Social Movement," *American Sociological Review*, 20 (1955), 3–10.

In the process of bargaining over side payments, many of the organizational objectives are defined or clarified. Because of the form the bargaining takes, the objectives tend to have several important attributes:

1. They are imperfectly rationalized. The extent to which the new demands will be tested for consistency with existing policy will depend on the skill of the leaders involved, the sequence of demands leading to the new bargaining, the aggressiveness of various parts of the organization, and the scarcity of resources. This testing is normally far from complete.

2. Some objectives are stated in the form of aspiration-level constraints. This occurs when demands that are consistent with the coalition are thus stated; for example: "We must allocate ten per cent of our total budget to research."

3. Some objectives are stated in a nonoperational form. In our formulation such objectives arise when potential coalition members have demands that are nonoperational or demands that can be made nonoperational. The prevalence of objectives in this form can be partly explained by the fact that nonoperational objectives are consistent with virtually any set of objectives.

3.2.2 Stabilization and elaboration of objectives

The bargaining process goes on more or less continuously, turning out a long series of commitments — but a description of goal formation in such terms alone is not adequate. Organizational objectives are, first of all, much more stable than would be suggested by such a model, and second, such a model does not handle satisfactorily the elaboration and clarification of goals through day-to-day bargaining.

Human beings have limited capacities and limited time to devote to any particular aspect of the organizational system; such limitations constrain the bargaining process. Let us return to our conception of a coalition having monetary r[illegible] policy side payments. These side-payment agreements are incomplete[illegible]ey do not anticipate effectively all possible future situations,

and they do not identify all considerations that might be viewed as important by the coalition members at some future time. Nevertheless, the coalition members are motivated to operate under the agreements and to develop some mutual control-systems for enforcing them.

One such mutual control-system in many organizations is the budget. ✓ A budget is an explicit elaboration of previous commitments. Although it is usually viewed as an asymmetric control-device (i.e., a means for superiors to control subordinates), it is clear that it represents a form of mutual control — just as there are usually severe costs to the department in exceeding the budget, so also are there severe costs to other members of the coalition if the budget is not paid in full. As a result, budgets in every organization tend to be self-confirming (see Chapter 5).

A second major, mutual control-system is the allocation of functions. ✓ Division of labor and specialization are commonly treated in management textbooks simply as techniques of rational organization. If, however, we consider the allocation of functions in much the way we would normally view the allocation of resources during budgeting, a somewhat different picture emerges. When we define the limits of discretion, we constrain the individual or subgroup from acting outside those limits; but at the same time, we constrain any other members of the coalition from prohibiting action within those limits. Like the allocation of resources in a budget, the allocation of discretion in an organization chart is largely self-confirming.

The secondary bargaining involved in such mutual control-systems serves to elaborate and revise the coalition agreements made on entry.[8] In the early life of an organization, or after some exceptionally drastic organizational upheaval, this elaboration occurs in a context where relatively deliberate action must be taken on everything from pricing policy to paper clip policy. Reports from individuals who have lived through such early stages emphasize the lack of structure that typifies settings for day-to-day decisions.[9]

In most organizations most of the time, however, the elaboration of objectives occurs within much tighter constraints. Much of the structure is taken as given. This is true primarily because organizations have memories in the form of precedents, and individuals in the coalition are strongly motivated to accept the precedents as binding. Whether precedents are formalized in the shape of an official standard operating procedure or are less formally stored, they remove from conscious consideration many agreements, decisions, and commitments that might well be subject to renegotiation in an organization without a memory (see Chapter 5). Past bargains become precedents for present situations; a budget becomes a precedent for future budgets; an allocation of functions becomes a precedent for future allocations. Through all the well-known mechanisms, the coalition agreements of today are institutionalized into semipermanent arrangements. A number of administrative aphorisms come to mind: an unfilled position disappears; see an empty office and fill it up; there is nothing temporary under the sun. As a result of organizational precedents, objectives exhibit much greater stability than would typify a pure bargaining situation. The "accidents" of organizational genealogy tend to be perpetuated.

---

[8] J. D. Thompson and W. J. McEwen, "Organizational Goals and Environment: Goal-setting as an Interactive Process," *American Sociological Review*, 23 (1958), 23–31.

[9] H. A. Simon, "The Birth of an Organization: The Economic Cooperat[illegible] lministration," *Public Administration Review*, 13 (1953), 227–236.

3.2.3 Changes in objectives through experience

Although considerably stabilized by internal processes, the demands made on the coalition by individual members do change with experience. Both the nature of the demands and their quantitative level vary over time.

Since many of the requirements specified by individual participants are in the form of attainable goals rather than general maximizing constraints, objectives are subject to the usual phenomena associated with aspiration levels. As an approximation to the aspiration-level model, we can take the following set of propositions:

1. In the steady state, aspiration level exceeds achievement by a small amount.
2. Where achievement increases at an increasing rate, aspiration level will exhibit short-run lags behind achievement.
3. Where achievement decreases, aspiration level will be above achievement.

These propositions derive from a set of assumptions requiring that current aspiration be an optimistic extrapolation of past achievement and past aspiration. Although such assumptions are sometimes inappropriate, the model seems to be consistent with a wide range of human goal-setting behavior.[10] Two kinds of achievement are, of course, important. The first is the achievement of the participant himself. The second is the achievement of others in his reference group.[11]

Because of these phenomena, our theory of organizational objectives must allow for drift in the demands of members of the organization. No one doubts that aspirations with respect to monetary compensation vary substantially as a function of payments received — so do aspirations regarding advertising budget, quality of product, volume of sales, product mix, and capital investment. Obviously, until we know a great deal more than we now do about the parameters of the relation between achievement and aspiration, we can make only relatively weak predictions. However, some of these predictions are quite useful, especially in conjunction with search theory (see Chapter 4).

The nature of the demands also changes with experience in another way. We do not conceive that individual members of the coalition will have a simple listing of demands, with only the quantitative values changing over time. Instead we imagine each member as having a rather disorganized file case full of demands. At any point in time, the member attends to only a rather small subset of his demands, the number and variety depending again on the extent of his involvement in the organization and on the demands of the other commitments on his attention.

Since not all demands receive attention at the same time, one important part of the theory of organizational objectives is to predict when particular units in the organization will attend to particular goals. Consider the safety goal in a large corporation. For the safety engineers this is an important goal most of the time. Other parts of the organization rarely even consider it. If, however, the organization has some drastic experience (e.g., a multiple fatality), attention to a safety goal is much more widespread and safety action quite probable (see Chapter 4).

Whatever the experience, it shifts the attention focus. In some cases (as in the safety example), adverse experience suggests a problem area to be attacked. In others, solutions to problems stimulate attention to a particular goal. An organization with an active personnel-research department will devote substantial attention to personnel goals, not because it is necessarily an especially pressing problem but because the subunit keeps generating solutions that remind other members of the organization of a particular set of objectives they profess.

The notion of attention focus suggests one reason why organizations are successful in surviving with a large set of unrationalized goals. They rarely see the conflicting objectives simultaneously. For example, consider the case of a common pair of organizational demands within business organizations, the demands for: (1) specific tailoring of product specifications and delivery times to individual customer needs — primarily from the sales department and customers; (2) product standardization and delivery times consistent with production smoothing — primarily from the production department and cost analysts. In large part, these demands are logically inconsistent; one is satisfied at the expense of the other. They cannot both be completely satisfied simultaneously. However, since the probability is low that both of these demands will be made simultaneously, the organization can remain viable by attending to the demands sequentially.

---

[10] K. Lewin, T. Dembo, L. Festinger, and P. Sears, "Level of Aspiration," in *Personality and the Behavior Disorders*, ed. J. M. Hunt (New York: Ronald, 1944), vol. I.

[11] L. Festinger, "A Theory of Social Comparison Processes," *Human Relations*, 7 (1954), 117-140.

The sequential attention to goals is a simple mechanism. A consequence of the mechanism is that organizations ignore many conditions that outside observers see as direct contradictions. They are contradictions only if we imagine a well-established, joint preference ordering or omniscient bargaining. Neither condition exists in an organization. If we assume that attention to goals is limited, we can explain the absence of any strong pressure to resolve apparent internal inconsistencies. This is not to argue that all conflicts involving objectives can be resolved in this way, but it is one important mechanism that deserves much more intensive study.

3.2.4 Organizational slack

In terms of the present framework, an organizational coalition is viable if the payments made to the various coalition members are adequate to keep them in the organization. If resources exist to meet all demands and those resources are distributed so as to meet demands, the coalition is a feasible one. Since demands adjust to actual payments and alternatives external to the organization, there is a long-run tendency for payments and demands to be equal. In this sense, what we have called *coalition demands* are analogous to the factor prices of a more conventional view of the firm.

There is a critical difference, however. In the present theory we focus on the short-run relation between payments and demands and on the imperfections in factor markets. The imperfections, in fact, dominate the behavior. The imperfections are dominant for three primary reasons:

1. As we have already noted, payments and demands are in the form of a variety of money payments, perquisites, policies, personal treatments, and private commitments. As a result, information on actual factor "prices" is hard to obtain, easily misinterpreted, and often unreliable.
2. Information about the "market" is not obtained automatically; it must be sought. Typically, the participants in the organization do not seek information until stimulated to do so by some indication of failure.
3. Adaptations in demands are slow — even in the face of strong pressure.

Because of these frictions in the mutual adjustment of payments and demands, there is ordinarily a disparity between the resources available to the organization and the payments required to maintain the coalition. This difference between total resources and total necessary payments is what we have called *organizational slack*. Slack consists in payments to members of the coalition in excess of what is required to maintain the organization. Many interesting phenomena within the firm occur because slack is typically [illegible] ero.

In conventional economic theory slack is zero (at least at equilibrium). In treatments of managerial economics, attention is ordinarily focused on only one part of slack — payments to owners — and it is assumed that other slack is maintained at zero. Neither view is an especially accurate portrayal of an actual firm. Many forms of slack typically exist: stockholders are paid dividends in excess of those required to keep stockholders (or banks) within the organization; prices are set lower than necessary to maintain adequate income from buyers; wages in excess of those required to maintain labor are paid; executives are provided with services and personal luxuries in excess of those required to keep them; subunits are permitted to grow without real concern for the relation between additional payments and additional revenue; public services are provided in excess of those required.

From time to time virtually every participant in any organization obtains slack payments. However, some participants ordinarily obtain a greater share of the slack than do other participants. In general, we would expect that members of the coalition who are full-time, in a position to perceive potential slack early, or have some flexibility in unilateral allocation of resources will tend to accumulate more slack than will other members. Although we have not attempted to pursue in detail the implications of such differential distribution of slack within the organization, the general model presented in Chapter 8 distinguishes between two kinds of organizational slack (essentially slack in the marketing-sales functions on the one hand and slack in the production function on the other) that are related by their mutual dependence on external conditions but are partially independent.

In most cases we have used the organizational slack concept not to explain differential payments but as a hypothetical construct for explaining over-all organizational phenomena. In particular, it seems to be useful in dealing with the adjustment of firms to gross shifts in the external environment. For example, consider what happens when the rate of improvement in the environment is great enough so that it outruns the upward adjustment of aspirations. In a general way, this seems to be the situation that faces business firms during strong boom periods. When the environment outruns aspiration-level adjustment, the organization secures, or at least has the potential of securing, resources in excess of its demands. Some of these resources are simply not obtained — although they are available. Others are used to meet the revised demands of those members of the coalition whose demands adjust most rapidly — usually those most deeply involved in the organization. The excess resources would not be subject to general bargaining because they do not involve allocation in the face of scarcity.

When the environment becomes less favorable, organiza[illegible] slack represents a cushion. Resource scarcity brings on renewed bargaining and tends to cut heavily into the excess payments introduced during plush times. It does not necessarily mean that precisely those demands that grew abnormally during better days are pruned abnormally during poorer ones, but in general we would expect this to be approximately the case. More important, the cushion provided by organizational slack permits firms to survive in the face of adversity. Under the pressure of a failure (or impending failure) to meet some set of demands on the coalition, the organization discovers some previously unrecognized opportunities for increasing the total resources available. For example, M. W. Reder reports that after losses of about fifty million dollars for the first three quarters of 1946, the Ford Motor Company "announced that it had found methods of reducing operating costs (on a given volume of output) by about twenty million dollars per year."[13]

Organizational slack absorbs a substantial share of the potential variability in the firm's environment. As a result, it plays both a stabilizing and adaptive role. We have already noted that the demands of participants adjust to achievement. Aspiration-level adjustment, however, tends to be a relatively slow process — especially downward adjustment. If the only adaptive devices available to the organization were adjustments in aspirations of the members of the coalition, the system would be quite unstable in the face of an environment of even moderate fluctuation. Slack operates to stabilize the system in two ways: (1) by absorbing excess resources, it retards upward adjustment of aspirations during relatively good times; (2) by providing a pool of emergency resources, it permits aspirations to be maintained (and achieved) during relatively bad times.

This is not to argue that slack is deliberately created for such a stabilizing purpose; in fact, it is not. Slack arises from the bargaining and decision process we have described, without conscious intent on the part of the coalition members to provide stability to the organization. In a sense, the process is reinforced because it "works" and it "works" partly because it generates slack, but we have seen no significant evidence for the conscious rationalization of slack in business firms. From the point of view of a behavioral theory of the firm, however, the critical question is whether predictions based on the concept can be verified. For example, we would predict that the costs of firms that are successful in the market place will, *ceteris paribus*, tend to rise. Such predictions are susceptible to more or less direct test.[14] They also are tested by testing more complicated models of which they form a part (see Chapter 5).

---

[13] M. W. Reder, "A Reconsideration of Marginal Productivity Theory," *Journal of Political Economy*, 55 (1947), 450–458.

[14] R. M. Cyert and J. G. March, "Organizational Factors in the Theory of Oligopoly," *Quarterly Journal of Economics*, 70 (1956), 44–46.

3.2.5 Constructing a predictive theory

Before the general considerations outlined above can be transformed into a useful predictive theory, a considerable amount of precision must be added. The introduction of precision depends, in turn, on the future success of research into the process of coalition formation. Nevertheless, some steps can be taken now to develop the theory. In particular, we can specify a general framework for a theory and indicate its needs for further development.

We assume a set of coalition members, actual or potential. Whether these members are individuals or groups of individuals is unimportant. Some of the possible subsets drawn from this set are viable coalitions. That is, we will identify a class of combinations of members such that any of these combinations meet the minimal standards imposed by the external environment of the organization. Patently, therefore, the composition of the viable set of coalitions will depend on environmental conditions.

For each of the potential coalition members we require a set of demands. Each such individual set is partitioned into an active part currently attended to and an inactive part currently ignored. Each demand can be characterized by two factors: (1) its marginal resource requirements, given the demands of all possible other combinations of demands from potential coalition members; (2) its marginal consistency with all possible combinations of demands from potential coalition members.

For each potential coalition member we also require a set of problems, partitioned similarly into an active and an inactive part.

This provides us with the framework of the theory. In addition, we need five basic mechanisms:

1. A mechanism that changes the quantitative value of the demands over time. In our formulation, this becomes a version of the basic aspiration-level and mutual control theory outlined earlier.

2. An attention-focus mechanism that transfers demands among the three possible states: active set, inactive set, not-considered set. We have said that some organizational participants will attend to more demands than other participants and that for all participants some demands will be considered at one time and others at other times, but we know rather little about the actual mechanisms that control this attention factor.

3. A similar attention-focus mechanism for problems. As we have noted, there is a major interaction between what problems and what demands are attended to, but research is also badly needed in this area.

4. A demand-evaluation procedure that is consistent with the limited capacities of human beings. Such a procedure must specify how demands are checked for consistency and for their resource demands. Presumably

such a mechanism will depend heavily on a rule that much of the problem be taken as given and only incremental changes considered.

5. A mechanism for choosing among the potentially viable coalitions. In our judgment, this mechanism will probably look much like the recent suggestions of game theorists that only small changes be evaluated at a time.[14]

Given these five mechanisms and some way of expressing environmental resources, we can describe a process for the determination of objectives in an organization that will exhibit the important attributes of organizational goal determination. At the moment, we can approximate some of the required functions. For example, it has been possible to introduce into several models substantial parts of the first four mechanisms (see Chapters 5, 7, 8, and 9).

3.3 Business Goals and Price and Output Decisions

Suppose we wish to use the general considerations noted above to construct a model of organizational decision making by a business firm determining price, output, and general sales strategy. As we have already noted, we are not yet in a good position to develop a theory that focuses intensively on the formation of objectives through bargaining and coalition formation (rather than on the revision of such objectives and selective attention to them). As a result, when we look at price and output determination in business firms, we do three things:

1. We assume a small set of operational goals. In making such an assumption we suggest that the demands of many parts of the coalition are not operative for this class of decisions most of the time or are substantially satisfied when the set of goals assumed is satisfied.

2. We assume that this set of goals is fixed in the sense that no other classes of goals will arise within the coalition. Such an assumption does not exclude changes in the levels of the goals nor in the attention directed at specific goals within the set.

3. We attempt to determine by empirical investigation what specific goals ordinarily enter into the price and output decisions. In general, we have observed that we can represent organizational goals reasonably well by using about five different goals. In any organization, other considerations sometimes arise. For example, governmental demands occasionally become of prime importance. In a few organizations other considerations are as important as those we have identified. For example, in some organizations considerations of prestige or tradition are major goal factors. However, for most price, output, and general sales strategy decisions in most organizations, we think we can limit our primary attention to five goals.

---

[14] D. L[illegible]d H. Raiffa, *op. cit.*

We list the five goals here in an arbitrary order without attempting to establish any necessary order of importance; most of the time no order of importance is required. All goals must be satisfied. However, it should be clear in the models we will present in later chapters that there is an implicit order in the models reflected in the way in which search activity takes place and in the speed and circumstances of goal-level change. These latent priorities appear to vary from organization to organization in a way that is not clear. It seems most probable that their variation should be explained in terms of differences in the bargaining position of the several participants in the coalition either current or historical, but at present we treat the implicit priorities simply as organizational parameters.

3.3.1 Production goal

We assume that an organization has a complex of goals surrounding the production operation. These can be summarized in terms of a production goal. Such a goal has two major components. The first is a smoothing goal: we do not want production to vary more than a certain amount from one time period to another. The second is a level-of-production goal: we want to equal or exceed a certain production level. These two components can be summarized in terms of a production range: we want production to fall within a range of possible production.

The production goal represents in large part the demands of those coalition members connected with production. It reflects pressures toward such things as stable employment, ease of scheduling, development of acceptable cost performance, and growth. Thus, the goal is most frequently evoked in the production part of the organization and is most relevant to decisions (e.g., output) made in that part.

3.3.2 Inventory goal

We assume certain aspirations with respect to finished-goods inventory levels. As in the case of the production goal, the inventory goal summarizes a number of pressures, most conspicuously the demands of some participants to avoid runouts in inventory and to provide a complete, convenient source of inventoried materials. We summarize these demands in terms of either an absolute level of inventory goal or an inventory range (in which case we also attend to demands to avoid excessive inventory costs).

The inventory goal reflects the demands of those coalition members connected with inventory. Primarily, thus, it builds on the pr[illegible]es of the inventory department itself, salesmen, and customers. Since t[illegible]ventory serves essentially as a buffer between production and sales, the goal is evoked most frequently and is most relevant to decisions in the output and sales areas.

3.3.3 Sales goal

We assume that most participants in business firms believe the firm must sell produced goods in order to survive. Thus, various members of the coalition make demands that the organization meet some general criteria of sales effectiveness. The sales goal and the market share goal (below) summarize these demands. In addition, the sales department itself (and the personnel in it) link subunit goals with sales. The sales goal is simply an aspiration with respect to the level of sales. It may be in terms of dollars, units, or both.

The sales goal represents primarily the demands of those members of the coalition closely connected with sales and secondarily those members of the coalition who view sales as necessary for the stability of the organization. The goal is most frequently evoked and is most relevant to decisions with respect to sales strategy.

3.3.4 Market share goal

The market share goal is an alternative to the sales goal insofar as the concern is for a measure of sales effectiveness. Either or both may be used, depending on the past experience of the firm and the traditions of the industry. In addition, the market share goal is linked to the demands of those parts of the organization that are primarily interested in comparative success (e.g., top management, especially top sales management) and to the demands for growth.

Like the sales goal, the market share goal is most frequently evoked and most relevant to sales strategy decisions.

3.3.5 Profit goal

We assume that the business firm has a profit goal. This goal is linked to standard accounting procedures for determining profit and loss. It summarizes the demands for two things: (1) demands for accumulating resources in order to distribute them in the form of capital investments, dividends to stockholders, payments to creditors, or increased budgets to subunits; (2) demands on the part of top management for favorable performance measures. In general, we assume that the profit goal is in terms of an aspiration level with respect to the dollar amount of profit. In principle, of course, this goal might also take the form of profit share or return on investment.

The profit goal reflects the pressure of those parts of the coalition that share in the distribution of profits and in the distribution of credit for profitability. Thus, in general, this pressure comes from top-level managers throughout the firm, from stockholders, creditors, and from those parts of the organization seeking capital investment. The goal is usually most closely linked to pricing and resource allocation decisions.

Although such a specification of goals deviates substantially from the conventional theory of the firm, it will not necessarily satisfy anyone who would like to reflect all of the goals that might conceivably be of relevance to price, output, and sales strategy decisions. Without insisting on the necessary efficacy of five goals, we think a strong case can be made for expanding the set of goals beyond that represented by the conventional theory, and even beyond the elaboration suggested by Baumol.[16] We also think that expanding the list of assumed goals much beyond the present list rapidly meets a point of diminishing returns. In the models presented in later chapters we restrict attention to this list of goals; in some cases a subset of goals seems satisfactory.

3.4 Summary

We have argued that the goals of a business firm are a series of more or less independent constraints imposed on the organization through a process of bargaining among potential coalition members and elaborated over time in response to short-run pressures. Goals arise in such a form because the firm is, in fact, a coalition of participants with disparate demands, changing foci of attention, and limited ability to attend to all organizational problems simultaneously.

In the long run, studies of the goals of a business firm must reflect the adaptation of goals to changes in the coalition structure. Except for some dramatic shifts, however, such changes are quite gradual, and it is possible to construct reasonable, short-run models in which a few specific types of goals are taken as given (subject to aspiration-level changes). With respect to the contemporary firm — and price, output, and sales strategy decisions — we have argued that we can identify five major goals: production, inventory, sales, market share, and profit. These goals, or a subset of them, are postulated in the models presented in later chapters (see Chapters 5, 7, 8, 9, and 10).

Finally, we have argued that, because of the form of the goals and the way in which they are established, conflict is never fully resolved within an organization. Rather, the decentralization of decision making (and goal attention), the sequential attention to goals, and the adjustment in organizational slack permit the business firm to make decisions with inconsistent goals under many (and perhaps most) conditions.

---

[16] W. J. Baumol, *Business Behavior, Value and Growth* (New York: Macmillan, 1959).

This, in turn, means that large size can increase the magnitude of the funds he can accumulate to finance further expansion. The businessman must have amassed wealth to be an efficient wealth amasser. These relationships alone can motivate businessmen to work hard to expand the scale of their operations.[11] Later in this book it is shown how this observation can be used as the basis for a model that determines the firm's equilibrium rate of growth.

---

[11] It is true that it also motivates them to expand by merger but, especially under the operation of the antitrust laws, merger is not always possible. And, in any case, growth by other means is also thereby encouraged.

We may also note, incidentally, that the profit pattern which has been described can make for growing inequality in the size distribution of firms and that the credit market can easily magnify this effect. It has been pointed out that a financial institution determines the amount it is willing to lend to any single borrower, at least in part, by estimating the prospective profitability of the investments for which the money is to be used. (*Cf.*, e.g., F. W. Taussig, *Principles of Economics*, New York, Macmillan, 3rd ed., 1927, p. 352.) But, in many cases, lenders base their estimate of prospective profitability on the profits earned by the borrower in the past. With time, this will only tend to increase the disparity between the amounts of money available to different entrepreneurs, for if my argument is correct, it is the entrepreneur with the greatest capital at his disposal who will have the largest present earnings and who can, therefore, most rapidly augment his funds by borrowing, as well as by accumulation.

CHAPTER 6

The revenue maximization hypothesis

Though businessmen are interested in the scale of their operations partly because they see some connection between scale and profits, I think management's concern with the level of sales goes considerably further. In my dealings with them I have been struck with the importance the oligopolistic enterprises attach to the value of their sales. A small reversal in an upward sales trend that can quite reasonably be dismissed as a random movement sometimes leads to a major review of the concern's selling and production methods, its product lines, and even its internal organizational structure.[1]

Before going on I must make an important terminological point. In ordinary business parlance the term "sales" refers not to the number of physical units of one of its products that has been sold but, rather, to the *total revenue* obtained by the firm from the purchases of its customers. In the near universal multiproduct firm, any measure of over-all physical volume must involve index number problems, and the adoption of a value measure doubtless is to be expected. In any event, in the sequel I shall adhere to the businessman's practice and use the terms "sales" and "total revenue" as synonyms. As a reminder, however, I will frequently employ "dollar sales" or "sales revenue" or some other such expressions.

1. Disadvantages of declining sales

There are many reasons why the businessman should show such concern about the magnitude of his sales. Declining sales can bring with them all sorts of disadvantages: There is reason to fear that consumers will shun a product if they feel it is falling in popularity, though their information on these matters is certainly often spotty. Banks and the money market will tend to be less receptive to the desires of a firm whose absolute or relative sales volume is declining. Perhaps even more important in this connection is the very real danger that firms whose sales are falling will lose distributors—a major marketing setback. Management also is not unmoved by the fact that in a declining firm personnel relations are made much more difficult when firing rather than hiring is the order of the day. The firm that declines (or remains small when others expand) can lose monopoly power and the power to adopt an effective competitive counter strategy when one is called for, and it may become more vulnerable to a general deterioration in business conditions. For all these reasons the executive may reasonably conclude that maintenance of as large a sales volume as possible is the only way to succeed in business.

---

[1] As has frequently been noted in the empirically based portion of the literature, the oligopolist seems to be equally fearful of any decline in the proportion between his sales and the sales of the rest of the industry. If its market share falls because its rivals are expanding more rapidly, the firm which has not kept up often reacts as though its sales were actually falling. For the moment I shall largely ignore this concern with market share because I deal first with the static theory. However, the point will play a more important role in the dynamic analysis of Part II.

Even if size did not promote profits, personal self-interest could well induce the managers of a firm to seek to maximize sales. Executive salaries appear to be far more closely correlated with the scale of operations of the firm than with its profitability.[2] And in the modern corporation, which is characterized so often by separation of ownership from management, many executives find it prudent to avoid an absolute or relative decline in their operations. Here, management's concern with the volume of sales is compounded of its very conscientious concern with the responsibilities of its trusteeship and a desire to play good stockholder politics. In any event, the effects are the same—the volume of sales approaches the status of a prime business objective.

2. Sales as an ultimate objective

Up to this point, in essence, I have been arguing that the firm may be expected to promote sales as a means to further its other objectives—operational efficiency and, ultimately, profits. So far, there is no necessary clash with orthodox analysis.

But now I propose to take the next step and suggest that the businessman has gone still further in his regard for sales volume. I believe that to him sales have become an end in and of themselves.

It must be made clear to begin with that this hypothesis in no way conflicts with an assumption of rationality. People's objectives are

---

[2] A statistical test of this surmise was undertaken by J. W. McGuire, J. S. Y. Chiu, and A. O. Elbing. See their "Executive Incomes, Sales and Profits," *American Economic Review*, Vol. LII, September 1962. Their results do seem to support the view in the text.

whatever they are. Irrationality surely must be defined to consist in decision patterns that make it more difficult to attain one's own ends, and not in choosing ends that, for some reason, are considered to be wrong. Unless we are prepared to determine other people's values, or unless they pursue incompatible objectives, we must class behavior as rational if it efficiently pursues whatever goals happen to have been chosen.

The evidence for my hypothesis that sales volume ranks ahead of profits as the main object of the oligopolist's concern is again highly impressionistic; but I believe it is quite strong. Surely it is common experience that when one asks an executive "How's business?" he will answer that his *sales* have been increasing (or decreasing) and talk about his profit only as an afterthought, if at all. And I am told the requirements for acceptance to membership in the Young Presidents Organization (an honorific society) are that the applicant be under 40 years of age and president of a company whose annual volume is over a million dollars. Presumably it makes no difference if this firm is in imminent danger of bankruptcy.

Nor is this failure to emphasize profits a pure rationalization or a mere matter of careless phrasing. Almost every time I have come across a case of conflict between profits and sales, the businessmen with whom I worked left little doubt as to where their hearts lay. It is not unusual to find a profitable firm in which some segment of its sales can be shown to be highly unprofitable. For example, I have encountered several firms that were losing money on their sales in markets quite distant from the plant, where local competition forced the product price down to a level that did not cover transportation costs. Another case was that of a watch distributor whose sales to small retailers in sparsely settled districts were so few and far between that the salesmen's wages were not made up by the total revenues they brought in. When such a case is pointed out to management, it is usually quite reluctant to abandon its unprofitable markets. Businessmen may consider seriously proposals which promise to put these sales on a profitable basis. There may be some hope for the adoption of a suggestion that a new plant be built nearer the market to which current transportation costs are too high or that watch salesmen be transferred to markets with greater sales potential and a mail order selling system be substituted for direct selling in little populated regions.[3] But a program that explicitly proposes any cut in sales volume, whatever the profit considerations, is likely to meet a cold reception. In many cases firms finally do perform the radical surgery involved in cutting out an unprofitable line or territory, but this usually occurs after much heart searching and delay.[4]

---

[3] Even this suggestion was not adopted by the watch distibutor.

3. The role of profits

It is tempting to object that along this road lies bankruptcy; and so it would if management were prepared not only to subordinate profit considerations to sales, but to disregard profits altogether. After all, maximum sales might require prices so low that the costs would nowhere be covered.[5] It is quite true that there is some conflict between the firm's sales goal and its profit objectives and, as is to be expected, the matter is settled by compromise. The compromise is, of course, usually tacit, its terms are ill-defined, and doubtless it varies from case to case. But I think it is nevertheless possible to set up a formal relationship which is analytically useful and, at the same time, provides us with a reasonably close approximation to the facts.

The nature of this approximation is again best suggested by an illustrative experience. A manufacturer of a new synthetic yarn indicated that he was reluctant to promote sales by introducing his product at a price that would not cover the cost of his small initial outputs. The firm's usual rate of return on investment played an explicit and very fundamental role in these deliberations. It was made clear that management was not concerned with obtaining profits higher than this. Once this minimum profit level was achieved, sales revenues rather than profits became the overriding objective.

I suspect that the much publicized practice of average cost pricing is a crude attempt to achieve just this sort of goal. Prices are set at a fixed markup above average costs, not only because this is a convenient rule of thumb, but also because the practice appears to set a floor under the rate of return.[6] Of course, it does not always work out in that way

---

[4] Yet, as will be shown in the next chapter, a good marginalist approach even to *sales* revenue maximization requires abandonment of disproportionately *unprofitable* sales segments. The firm's resources should be allocated among its different products, territories, etc. in such a way that the marginal *profits* from all types of sales are equal. For if the sacrifice of a dollar of profit can produce a greater addition to sales of type A than to sales of type B, then sales cannot be at a maximum (subject to the profit constraint described in the next section of the text) and effort should be transferred from the promotion of B sales to the promotion of sales of A.

[5] But note that maximization of *dollar* sales will never call for zero or negative prices because, at these prices, sales (equal total revenue) must also be zero or negative.

[6] It also helps to prevent profits from going too high and may thereby discourage the entry of new competitors. This consideration explains, in part, the firm's failure to seek to maximize short-run profits. However, I feel that its role in the real firm's calculations has been exaggerated a bit in some of the attempts to rationalize average cost pricing.

because volume can be miscalculated and cost estimates may therefore turn out to be incorrect. But the objective of the procedure seems clear nevertheless.

I am prepared to generalize from these observations and assert that the typical oligopolist's objectives can usefully be characterized, approximately, as sales maximization subject to a minimum profit constraint.[7] Doubtless this premise overspecifies a rather vague set of attitudes, but I believe it is not too far from the truth. So long as profits are high enough to keep stockholders satisfied and contribute adequately to the financing of company growth, management will bend its efforts to the augmentation of sales revenues rather than to further increases in profits.

4. Determination of the minimum profit level

We have yet to discuss how the minimum acceptable profit level is determined. Once we have taken care of this loose end we shall have laid the foundation for a self-contained static oligopoly model, as will be shown in the next chapter. But until something is said about the determination of the minimum profit level the entire analysis is left up in the air. So long as the acceptable profit level is left undetermined, almost any price or output decision is compatible with the stated objectives of the firm.

Once more, almost anything that is said here must be an oversimplification and an overrigid characterization of a fairly amorphous and variable set of business practices. Yet I believe that the following discussion does not represent a very great distortion of the facts.[8]

Like the profit maximizer, the sales maximizing firm must think of tomorrow. Its horizon will normally extend well into the future and, accordingly, it must consider prospective sales along with the sales level it is currently achieving. Then increased future sales also will be encompassed in the basic sales maximization objective.

For its outputs and its sales to grow, the firm will need funds to finance its expansion. These funds can be obtained either internally, out of retained profits, or from external sources—by borrowing or through the issue of new equity. But the availability of external funds to the firm will depend also on its earnings.[9] The larger the earnings of the company, the more financing it will be able to obtain from the capital market. The firm will then find it desirable to keep up its profit level as a means of financing future sales—but at the same time, high profits, as we shall see, are likely to require some limitation in its current sales volume. Thus, higher profits promote the firm's goals in one way and conflict with them in another. From this long-run point of view, there is then likely to be some intermediate level of profits that is optimal—one that most effectively serves management's sales maximization objective. Looked at in this way the profit level does not take the form of a constraint. Rather, it becomes an instrument variable whose value is determined as part of the optimality calculation. In Chapter 10, when we get to our company growth models, this will be seen very clearly.

---

[7] There is nothing new in the observation that firms seek to meet a minimum profit requirement. Thus, e.g., Professor Lintner writes, "The available evidence indicates that many if not most managements do have in mind some sort of minimum return which contemplated investments must satisfy if they are to be undertaken. . . ." (John Lintner, "Effect of Corporate Taxation on Real Investment," *American Economic Review*, Vol. XLIV, May 1954, p. 522.)

[8] The argument which follows is based on a suggestion by Professor Leibenstein, for which I am profoundly grateful.

But in the short run the required profit appears to be given exogenously. It does serve as a constraint circumscribing the firm's decisions—one which requires it to provide a rate of return to stockholders sufficient to make certain that the company's securities remain attractive to the capital market.[10]

5. Afterthoughts on sales maximization

Since the publication of the first edition of this book the theoretical literature on managerial motivations has blossomed forth and provided a number of alternative and plausible surmises on the aims of business

management. Among these suggestions have been a model in which the size of staff enters the managerial utility function,[11] one in which the firm wishes to maximize profit subject to a constraint on sales,[12] and one in which growth becomes a primary company objective.[13] Of course, the notion that the firm may have objectives other than profits antedates considerably the appearance of this volume.[14] However, the distinguishing characteristic of the more recent writings is the fact that they provide analytical models—that they are operational in the sense that they can yield theorems as interesting as those that can be derived from a profit maximization approach.

I have no basic quarrel with any of these, for it seems to me that any contribution provided by the models of this book consists primarily in the demonstration that alternative assumptions about managerial motivations are tractable analytically. I have never claimed that there is any "correct" model that captures the whole truth about managerial goals in all their subtle shadings and complexity. As a matter of fact, diversity of objectives and variation in motives with the passage of time naturally is to be expected if I am right in my view that it is no more rational for managements to hold one consistent set of goals than it is for them to pursue any other.

It is for this reason that I have been surprised at how consistently the firms with which I have had dealings at least appear to have held to some sort of sales and growth objectives. I have come across only one very clear exception, in which the new president of a major company was quoted in the press to the effect that he did not aim to make his firm the largest in the industry—only the most profitable. It is interesting that this man had once been an academic economist.[15]

---

[9] In a pure Miller-Modigliani world they may depend only on the firm's earnings and be totally unaffected by their allocation between dividends and retained profits. After all, retained earnings if properly reinvested yield increased future earnings which will bring capital gains to the stockholder. If the firm has made its investment plans independently of its financial decisions (surely optimal investment decisions are not dependent on the sources of the funds used to finance them), then an increase in dividend payments should make absolutely no difference one way or the other to current stockholders. With future plans fixed, the total current value of the firm, V, will also be fixed. Suppose there is a D dollar dividend that is made up for by a D dollar new stock issue. The market value of current stockholders' claims on the firm will have declined from V to $V - D$ which is exactly compensated by their D dollar dividend. See M. H. Miller and F. Modigliani, "Dividend Policy, Growth, and the Valuation of Shares," *Journal of Business*, Vol. XXXIV, October 1961.

[10] This argues, then, that from the point of view of the stockholder, management may be what Simon has called a "satisficer" rather than a maximizer—it may seek merely to provide a return sufficient to keep stockholders contented.

Incidentally, because bonds and stocks are substitutes both from the view of the buyer (lender), and from that of the borrower firm that issues them, it may also be argued that, in a rough way, dividends will in the long run also be tied, via the risk premium, to the level of interest rates. Though in the short run there is apparently little correlation between the earnings on stocks and the yields of fixed interest securities, competitive pressures, in the long run, must drive their earnings toward approximate parity, otherwise the type of security whose yield was consistently lower would eventually run out of takers. If management wishes to maintain dividends and stockholders' capital gains as determined by the course of the price of the company's securities at a level which will be considered adequate, the minimum acceptable rate of profit is then determined by the same apparatus that sets the levels of interest rates. In other words, oligopoly earnings net of risk premium would then approximate the so-called normal rate of profit—that rather mysterious ingredient in the economist's cost calculations.

[11] See O. E. Williamson, *The Economics of Discretionary Behavior*, Englewood Cliffs, Prentice-Hall, 1964, pp. 40ff. See also R. J. Monsen, Jr., and A. Downs, "A Theory of Large Managerial Firms," *The Journal of Political Economy*, Vol. LXXIII, No. 3, June 1965, in which it is assumed that "managers act so as to maximize their own lifetime incomes," p. 225.

[12] See Franklin M. Fisher's review of this book, *Journal of Political Economy*, Vol. LXVIII, June 1960, and D. K. Osborne, "On the Goals of the Firm," *Quarterly Journal of Economics*, Vol. LXXVIII, November 1964.

[13] E.g., Harvey Leibenstein, *Economic Theory and Organizational Analysis*, New York, Harper & Row, 1960, Part IV, and R. Marris, *The Economic Theory of Managerial Capitalism*, Glencoe, The Free Press, 1964, especially Chs. 2 and 3. See also Ch. 10 in this book.

[14] Thus, see the references in Williamson, *op. cit.*, Ch. 2. See also R. F. Lanzillotti, "Pricing Objectives in Large Companies," *American Economic Review*, Vol. XLVIII, December 1958.

I will not attempt to reply systematically to criticisms that have been leveled against the sales maximization hypothesis. After all, the question at issue is a factual matter that cannot be settled by *a priori* arguments. Yet it may be useful to comment on a few of these; for example, the suggestion that, for some reason that is to me obscure, companies that pursue this goal are unlikely to survive very long. Indeed, one noted economist caught himself about to argue that a profit maximizing firm would end up outselling its sales maximizing competitor! By definition, the (long-run) sales maximizer must end up the larger while the profit maximizer produces more income for its stockholders, but neither type is an inherently shaky proposition. If long-run sales objectives of firm is an inherently shaky proposition. If long-run sales objectives require larger profits, the rational sales maximizer will do all he can to obtain these profits as an intermediate step toward the achievement of his objective.

Another argument that has been raised against the sales maximization hypothesis asserts that all of the (rather miscellaneous) phenomena that have been used to defend and illustrate it are perfectly consistent with a *long-run* profit maximization goal. That is, the firm may seek to expand its sales today primarily because it considers this to be a good means to earn more money tomorrow. I would still guess that this alternative hypothesis is not generally valid, but I am really somewhat skeptical about the very concept of the long-run goal. It is particularly questionable in a large firm in which there are many centers of power of varying potency and where goals have never been discussed explicitly and formulated (except for purposes of public relations releases, in which the promotion of virtue itself, in effect, is described as the ultimate aim of the corporation). If observed behavior is perfectly consistent either with long-run sales or profit maximization goals, and we are to avoid its attribution to a corporate unconscious that has somehow adopted its goals without their having been revealed to management, must we not simply file suggestions for *long-run* objectives under the heading "not relevant," or at least "not settled"?

---

[15] Other exceptions may be extremely large firms, such as General Motors, whose managements fear that further expansion will arouse the interest of the trust busters. However, these giants also frequently try to expand by increasing their product lines while holding their market share constant on older products. On the other hand, the mutual life insurance companies, for whom profits are not even readily defined, may be particularly obvious sales maximizers.

CHAPTER 7

A static oligopoly model

The empirical hypothesis of the preceding chapter about the characteristic objectives of modern oligopolistic firms can be used as the basis of a formal analysis that yields several interesting conclusions. Most of this can be accomplished with the aid of fairly simple geometry, but I shall employ some algebra at the end of the chapter in order to generalize a few of the propositions. Discussion of the main implications of the analysis has been deferred to later chapters.

1. Price-output determination: partial analysis

It will be remembered that *sales* maximization under a profit constraint does not mean an attempt to obtain the largest possible physical volume (which is hardly easy to define in the modern multiproduct firm). Rather, it refers to maximization of total revenue (dollar sales), which is to the businessman the obvious measure of the amount he has sold.[1] Maximum sales in this sense need not require very large physical outputs. To take an extreme case, at a zero price physical volume may be high but dollar sales volume will be zero. There normally will be[2] a well-determined output level that maximizes dollar sales. This level ordinarily can be fixed with the aid of the well-known rule: Maximum

---

[1] We may be tempted to say then that he suffers from money illusion. But this can hardly be taken as a criticism of his rationality—his objectives simply are what they are and there is nothing inherently rational or irrational about them.

[2] Provided there is some price sufficiently high to drive the quantity demanded down to zero, the total revenue function will be bounded from above, i.e., total revenue will be less than or equal to this maximum price multiplied by the quantity which will be demanded when price is zero. The total revenue will be zero both at this maximum demand price and when the price is zero. At intermediate prices it will be positive provided the demand curve has a negative slope throughout. Hence, in such cases at least one maximum total revenue point must exist. If the demand curve is everywhere differentiable such a point will only occur where the price elasticity of demand is unity.

1

INTRODUCTION

This book is about the business firm and the way it makes economic decisions. We propose to make detailed observations of the procedures by which firms make decisions and to use these observations as a basis for a theory of decision making within business organizations. Our articles of faith are simple. We believe that, in order to understand contemporary economic decision making, we need to supplement the study of market factors with an examination of the internal operation of the firm — to study the effects of organizational structure and conventional practice on the development of goals, the formation of expectations, and the execution of choices.

The rationale for such a belief is also simple. The modern "representative firm" is a large, complex organization. Its major functions are performed by different divisions more or less coordinated by a set of control procedures. It ordinarily produces many products, buys and sells in many different markets. Within the firm, information is generated and processed, decisions are made, results are evaluated, and procedures are changed. The external environment of the firm consists, in part, of other firms with comparable characteristics. If the market completely determined the firm's economic behavior, these internal attributes would be little more than irrelevant artifacts. But the market is neither so pervas or so straightforward. The modern firm has some control over the market; it has discretion within the market; it sees the market through an organizational filter.

The elaboration of these articles of faith into a theory of the firm involved four major research commitments. They are commitments that evolved during the course of the research, but they constitute a general retrospective characterization of our research strategy:

1. *Focus on a small number of key economic decisions made by the firm.* In the first instance, these were price and output decisions; subsequently they included internal allocation and market strategy decisions.

2. *Develop process-oriented models of the firm.* That is, we viewed decisions of the firm as the result of a well-defined sequence of behaviors in that firm; we wished to study the decisions by studying the process.

3. *Link models of the firm as closely as possible to empirical observations* of both the decision output and the process structure of actual business organizations. The models were to be both explicitly based on observations of firms and subject to empirical test against the actual behavior of identifiable firms.

4. *Develop a theory with generality beyond the specific firms studied.* We wanted a set of summary concepts and relations that could be used to understand the behavior of a variety of organizations in a variety of decision situations.

Each of these commitments was critical to the research; each in some way characterizes the theory that resulted.

The focus on a specific set of economic decisions (i.e., price and output) was intended to constrain the tendency of theories of decision making to become excessively general. Quite simply, we wanted models that could predict actual decisions (if provided with necessary data), not simply models that could predict some qualitative properties of those decisions. Obviously, we wanted both kinds of predictions if we could manage them, but by restricting our focus we hoped to strengthen the specificity of the models.

The emphasis on studying actual decision processes implies a description of the firm's decision in terms of a specific series of steps used to reach that decision. The process is specified by drawing a flow diagram and executing a computer program that simulates the process in some detail. We wanted to study the actual making of decisions and reproduce the behavior as fully as possible within the confines of theoretical manageability.

The insistence on empirical content follows naturally from the emphasis on identifying the actual process of decision. It is hard to imagine constructing a process model without detailed observation of the ways in which business organizations make decisions. Thus, the empirical commitment

has forced us to reconsider the traditional truculence of the firm with respect to outside investigation. We were constrained to secure research access to a number of firms under conditions that protected the competitive position of the firm without interfering with the disclosure needs of a research study.

Finally, the commitment to theoretical generality was intended to complement the insistence on specificity in models with a similar insistence on generality of concepts and broad structure. We did not want the theory to be limited to descriptive models of specific firms making specific decisions. We required a set of explanatory concepts that would serve as a framework for analysis and further research.

The results of our efforts to develop an empirically relevant, process-oriented, general theory of economic decision making by a business firm are summarized in this book. We hope they suggest both the utility and some of the problems of theoretical research on decision making in the firm. More generally, we hope they support the case for the legitimacy and importance of the business firm as a unit for detailed empirical study and theoretical analysis. More specifically, we hope that the rudiments of a behavioral theory of the firm presented in the following chapters will prove relevant both to economic theory and to the theory of complex organizations.

4-6

3

WHARTON REPROGRAPHICS

14

Journal of Economic Literature,
June 1979, Vol. XV

A Branch of Economics is Missing: Micro-Micro Theory

By HARVEY LEIBENSTEIN

Harvard University

and

Institute for Advanced Study
Princeton, New Jersey

I wish to thank James Dean, James Medoff, Richard Day, Victor Goldberg, Albert Hirschman, Gordon Winston, Persio Arida, and the editor for helpful comments.

I. *Introduction*

IN THE LIFE HISTORY of most sciences there are movements toward the study of larger aggregates or toward the detailed study of smaller and more fundamental units. My impression is that in most fields the movement toward the study of more micro units has predominated. Yet in economics in the 1930's the movement was in the macro direction. Both physics and biology (in the last three decades) have made great strides by studying smaller and smaller entities—physics by studying more minute fundamental particles and biology by studying the fundamental elements determining genetics. In a general sense economics has not been moving in this direction, although some work of this nature exists. The purpose of this paper is to review some samples of the work that exists and to argue that this area must become a major field of economic research and study.

The question of how individuals in multiperson firms influence firm decisions seems like such a natural question to ask that it is amazing that it is not part of the formal agenda of economists as a profession. Of course it has been asked, but not by present-day economists in their professional capacity. In other words, micro-microeconomics has never become an established field. For the most part, theorists have not only not raised this question, but they have continued to develop micro theory in such a way as to discourage economists from raising this question. Part of the reason for this lies in the maximizing and optimizing biases of conventional micro theory, and part can be ascribed to the consequences of the long period required to refine the theory so that elements that did not fit the basic model were discarded.[1]

---

[1] Both Alfred Marshall in *Industry and Trade* [33, 1923] and Frank Knight in *Risk, Uncertainty and Profit* [21, 1921] considered intra-firm elements. However these considerations have certainly dropped out of modern microeconomic analysis. They do not appear to be necessary elements of the main-line neoclassical model.

478 A frequently-used metaphor is that the firm behaves *as if* it possessed a "black box," which drives the firm *as if* it were an individual, irrespective of size. One way of looking at micro-micro theory is to suggest that the micro-micro problem is the study of what goes on *inside* the black box. Does the multiperson firm face the same problems as the one-man firm? What are the components of the black box? From this point of view, micro-micro theory is concerned with *intra firm* behavior and relations or with the interaction of persons within the firm and their influences on firm behavior. Outside the degenerate case of the one-man firm, for which the boundary between micro and micro-micro disappears, the distinction between the two types of theories, micro and micro-micro, is unambiguous.

H1

An obvious question is the relation between firm owners and firm members. Does the owner completely control, through contracts or other means, what others in the firm do? Can costs be understood without first answering this question? It would be platitudinous to argue that only to the extent that intraorganization activities are understood can costs of production be understood. But this is not a view that one gets from traditional micro theory, which simply asserts that costs are minimized. Hence, we have at least one reason why we should be interested in micro-micro theory. Another is whether the picture of the firm we get by analyzing the internals of the black box independently from conventional theory leads to a view of the firm consistent with conventional micro theory.

For the most part these questions are not answered in the conventional texts; and from this standpoint micro-micro theory may be said to be missing. However, there are other writings that come close to investigating some aspects of this problem. While these writings are too voluminous to review in detail, it is useful to review them on a broad brush basis in order to see whether a valid argument can be made that micro-micro theory as a branch of economics is missing, for the most part. Further, I will argue that this branch represents at this point a natural and useful research frontier.

To carry out this task we will examine, (1) complex objective functions, (2) the work of the Carnegie school, (3) what I have referred to elsewhere as X-efficiency theory, and (4) some related writings that do not fit easily into any of these categories.

II. *Complex Objective Functions versus Micro-Micro Theory*

There are at least two ways of trying to create alternatives to the traditional micro theory: One is to change the objective function of conventional theory, while the other is to start by creating micro-micro theories. These two approaches are not the same, although they are sometimes confused for the same thing. In this connection it is of interest to look at several well known "complex objective function" theories: those associated with the names Tibor Scitovsky, William J. Baumol, Robin Marris, and Oliver E. Williamson.

Scitovsky's theory appeared in the 1940's [44, 1943]—and his work may be looked at as the paradigm for the others. According to Scitovsky, the neoclassical assumption that an entrepreneur maximizes profits as well as satisfaction implies that the entrepreneur's choice between more income and more leisure is independent of his income. This presupposes a special hypothesis about the psychology of the entrepreneur such that his profit-seeking activities will continue unabated as his income increases. In contrast to the neoclassical entrepreneur, Scitovsky's entrepreneur trades income for leisure as his income rises. Whether or not entrepreneurs in general maximize profits can only be determined empirically. In this model there is no distinction between entrepreneurs, owners, and managers.

Baumol suggested an objective function that emphasized sales maximization subject to a profit constraint [5, 1959]. Baumol came to this conclusion as a result of his work with a business consulting firm. The profit constraint has something to do with the belief that the stockholders have to be satisfied in some sense. However, there is nothing in Baumol's theory that determines what the profit constraint ought to be. Nevertheless, for any given positive profit constraint, Baumol was able to demonstrate that he could get qualitive results different from the conventional profit maximizing model.

The growth maximization model developed by Robin Marris [30, 1963] is in some respects different from Baumol's, but is similar to it in spirit. Clearly there are similarities between growth and sales maximization, but they need not be the same thing. Thus, reinvestment is important in the Marris model. Furthermore, the Marris model finds a solution for the profit constraint; *i.e.,* profits must not be so low as to reduce the value of the stock to a degree that a takeover bid is probable. Presumably the manager's position is vulnerable to the introduction of new owners. Thus Marris's managers work out trade-offs between the utility of growth versus the utility of profits so that the position of management is not threatened.

Williamson [50, 1963] developed a model that has greater similarity to Scitovsky's than to Baumol's and Marris's. Williamson assumes that managers control the firm and are in a position to divert profits for the benefit of management—but only indirectly. Profits belong to the stockholders. Managers cannot cut into profits after the fact—that is, after they have been earned. But managers can lower profits by increasing expenses. Thus Williamson implicitly assumes that there exists a class of expenses, a subdivision of all expenses, which is for the benefit of managers. Some examples come easily to mind: lavish offices, carpets, extra secretarial services, entertainment budgets, and so on. Other "benefits" may be more subtle and be related to taking it easier than otherwise—which manifests itself in organization slack.

In any event, because of managerial discretion, expenses are higher and profits lower than would be the case under profit maximization. In this model we visualize the manager (or managers) deciding, as it were, on the distribution between profits and management perquisites so as to maximize the utility of a profits-perks mix.

If we combine all four of these models, we can come up with a composite objective function. Under such a function the managers would maximize the utility of a function whose arguments would include: profits, sales, the growth rate, and managerial perquisites. By changing the arguments, we can develop almost any number of different complex objective function models.

There are problems with such a vision of firm behavior. A basic one has to do with the distinction between a single decision-maker versus group decision-making. Do groups decide the same way as individuals do? Are the same ideas as meaningful for groups as they are for individuals? These are fundamental questions that permeate many aspects of our discussion. We shall treat them in this section only in connection with the problem of complex objective functions. Consider Williamson's function. Is the "Manager" who decides between profits and "enjoyable expenses" an individual or a group? If there is only *one* manager, than no problem arises. It makes sense for the manager to maximize a utility function. Such a person can decide between the utility of profit versus the utility of perquisites disguised as expenses. What meaning can we give to the notion of maximizing utility of a group of individuals? How is the decision made between different individuals' evaluation of the utility of more versus less profits?

The main point of all this is that the basic question of micro-micro theory is not answered by the complex objective function approach. It retains the black box of conventional micro theory, but changes the assertion about the behavior of the black box. These remarks are not intended as a criticism of Williamson or of any of the other authors involved. They did not set out to solve a micro-micro problem but to make micro theory more realistic. In this they succeeded. Nevertheless, these models do not handle the problems that arise when we shift from the single person dictator-manager case to the multiperson case. In general, it would seem reasonable to expect that (1) the larger the management group, the greater the number of arguments in the objective function and (2) the larger the potential perquisites to be distributed among managers, the greater the problem.

The essence of the problem is the size of the management group and the relation of various members to each other. For different individuals there are likely to be differences in power and similarity or conflicts of interests that arise. Many organizations allow for conflict resolution through a hierarchical system so that differences among lower status individuals can be resolved by higher status individuals. But despite a hierarchical power structure, informal degrees of influence may play a significant role. Hence, the multiperson situation as such creates difficulties in determining how "utilities" are distributed, and what meaning can be attached to utility maximization where actual or potential conflict exists. In other words, there is no reason to assume that any system of conflict resolution necessarily results in "joint maximization."

A single individual can be presumed to have a unique objective function, at least at any one time. But do two individuals necessarily have the same objective function? Whatever the answer, this is clearly a legitimate question. By adding enough arguments it may always be possible to presume that a utility function exists for any number of individuals. Thus if $U_a(X_1, X_2)$ and $U_b(X_3, X_4)$ are two objective functions for individuals *A* and *B*, a combined function $U(X_1, X_2, X_3, X_4)$ may be formed where each X_i stands for a different subobjective. But is all this meaningful if there are hundreds of X's? Clearly, the "team sense" of the management group must become attenuated as sub-objectives multiply. Does the sum of all the individual objectives, or some subset of these, add up in a way to a group objective? At least in some cases one can argue that there is a sense of group objectives that differs from the sum of individual objectives.

For a multiperson group such as a committee or a group of committees, maximization of utility may be an inappropriate concept. There is the game-theoretic problem made famous in the John von Neumann and Oskar Morgenstern's *Theory of Games* [38, 1944]. If an individual (or decision-maker) does not control all the variables, then maximization by any decision-maker may be impossible. This is one essential of the games of strategy problem. If *A* controls options X_1 and X_2, while *B* controls X_3 and X_4, the outcome depends on how each chooses. There is no necessary way for *A* or *B* to choose a maximum *individually*. Whether or not they can do so jointly (*i.e.*, cooperatively) is another matter. In the same sense, if management is a multiperson group, there may be no way to guarantee maximizing the utility of the group. Also, "forced" cooperation may also not yield a maximum. Suppose part of the value for some of the managers is noncooperation in some instances. Cooperation may reduce the utility of some individuals who enjoy an individualistic (non-cooperative) stance. The most likely outcomes are situations in which some gain and others lose. Thus a maximum in the Pareto superior sense may be ruled out in such instances. In sum, groups involve potential conflict, which puts into question certain basic concepts that are applicable to individuals.

III. *"Semi Micro-Micro Theory"— The Carnegie School*

The set of ideas developed by Herbert Simon, his colleagues, and those (many of whom were at one time or another at Carnegie-Mellon University) who were influenced by his ideas, will be referred to as the Carnegie school ideas. (See Kalman Cohen and Richard Cyert [9, 1975].)

In general, the writings of the Carnegie school contain an extremely rich set of ideas about the internal operations of firms. This point cannot be overemphasized. A major concern of many of these writers is organization theory and the analysis of a variety of intraorganizational problems. Thus, anyone searching for ideas in this area would be well advised to pursue the work of this group. At the same time, from the standpoint of our problem area there are difficulties with this body of literature. It has not resulted in an alternative to the existing neoclassical paradigm, although it may contain the raw materials for constructing one or more such alternatives. These remarks are not intended as a criticism of what has been achieved, but indicate the boundaries of the effort to date.

This set of ideas is rather difficult to place in terms of the dichotomy established for purposes of this paper. Most of the central ideas do not of themselves imply a micro-micro model. To start with, consider Simon's famous concept, *satisficing*. Satisficing behavior is an alternative to maximizing behavior. However, it could be applied just as easily to an individual decision-maker, or a group, or a multiperson firm. It substitutes a different decision criterion for maximization or optimization, but it does not of itself imply that we go inside the firm for this purpose.

The same can be said for a number of other concepts. Richard Day [12, 1964; 13, 1975] has attempted to summarize the essentials of the Carnegie approach in a terse and useful way for our purposes. He lists ten basic characteristics, one of which is (1) satisficing. Let us look briefly at the others and see to what extent they involve a micro-micro approach. These are: (2) bounded rationality, (3) multiplicity of goals, (4) sequential attention to goals, (5) feedback, (6) standard operating procedures, (7) resistance to change except under "duress," (8) coalitions to resolve conflicts, (9) organizational slack to stabilize coalitions, and (10) maintenance of viability. In Day's list, eight of the ten characteristics are not *necessarily* related to an essentially micro-micro approach. Most of them are related to the departure from the perfect rationality assumption. The two that are related to the *insides* of the firm are (8) "coalitions to resolve conflict," and (9) "organizational slack to stabilize coalitions." But we are not presented with a theory of intra-firm coalitions. That is, we are not told how coalitions are formed, who gets into a coalition, who remains outside, who gets what in a coalition, and so on. Of course, ideas not necessarily related to micro-micro models may be extremely useful in developing such models. Thus, the notions of (3) multiplicity of goals, (6) standard operating procedures, and (9) organizational slack lend themselves readily to the creation of micro-micro type simulation models. However, both the spirit and essence of the ideas, as far as I understand them, are that the firm itself is a coalition, and organizational slack exists to maintain that coalition.

I find it extremely difficult to judge the depth of the micro-micro concern of the

Carnegie school approach. Interpreters and summarizers have not emphasized that aspect. To the extent that emphasis exists, it is on the substitutes for the maximization assumption. Where the Carnegie school is contrasted with the neoclassical position, it is almost invariably the case that the deviations from rationality are emphasized. The elements usually discussed are satisficing, aspirations, bounded rationality, multiplicity of goals, standard operating procedures, and feedback.

An essential difficulty in treating the Carnegie school is that one cannot ask what the Carnegie economic model is and expect to be shown a specific article or book and to be told, "This is it"—at least, not in the sense in which a good graduate (or high level intermediate) text represents neoclassical theory or in which Gerard Debreu's book [15, 1959] or Kenneth Arrow and Frank Hahn [4, 1971] represents modern general equilibrium theory. From this point of view the text by Cohen and Cyert is of considerable interest [9, 1975]. The authors present both the standard neoclassical theory of the firm and some Carnegie school writings, as well as other alternatives. One of the authors, Cyert, has been a major contributor to the work of the Carnegie school and with James G. March wrote *The Behavioral Theory of the Firm* [11, 1963].

In their text Cohen and Cyert face the contrast between the Carnegie school and the orthodox approach. It is of interest that they step away from the possible conflict between the two types of theories and search for a conciliation between the two views. Some quotations from the Cohen and Cyert book are of interest:

> At this level of generality, we think that the results deduced from a set of assumptions which include utility and profit maximization are perfectly acceptable. If we wish to predict how a *particular* household or firm will behave, then assumptions this general will not be satisfactory. [9, 1975, p. 51.]

> Subjectively rational decision models of the type proposed by Simon represent a promising alternative to the usual models of economists which implicitly assume that great knowledge and computational ability are possessed by decision makers. Evidence suggests that Simon's approach to economic decision making is likely to provide good explanations and predictions of the actual decision-making behavior of consumers and businessmen. [9, 1975, p. 308.]

But earlier we read that

> The analysis of microeconomics assumes rational behavior by individual economic agents. The usefulness of this analysis for answering questions involving the allocation of resources among broad economic sectors, such as markets or industries, does not necessarily depend upon the extent to which consumers actually do maximize their utilities and business firms actually do maximize their profits. [9, 1975, p. 289.]

The view taken seems to be that for industry-wide or larger problems the neoclassical approach is appropriate, while those involving predictions of specific firm behavior may be aided by applying some of the newer theories. Obviously there is an unresolved procedural difficulty here. If a theory will not predict the behavior of a firm, can it predict the behavior of a collection of firms? The foregoing is not to suggest that there is no satisfactory answer to this question, but somewhere this question should be investigated. Perhaps Cohen and Cyert felt that their text was not the place to do so.

At the same time one gets the impression from the Cyert and Charles Hedrick paper [10, 1972, p. 409] that the authors share the "growing uneasiness with the neoclassical approach," and they conclude that "the problem is clearly difficult, but we wonder whether economics can remain an empirical science and continue to ignore the actual decision-making process of real firms."

Yet the view represented by the quotations may be consistent with the present status of Carnegie school theories. The impression that one gets is that there is no general Carnegie school model. Rather, what we have is a set of ingredients; we are asked to work out specific recipes of our own from these ingredients. Sample recipes already exist. These are the simulation models presented in Cyert and March's *The Behavioral Theory of the Firm* [11, 1963] and elsewhere. While these simulation models are examples of the approach, they need not be consistent with any more general Carnegie school theory. Each simulation model represents a single case. Since the bounds of the simulation model are not constrained by a more general theory, it would seem reasonable to assume that we cannot determine in advance whether the simulation model is consistent with the neoclassical model. (See also William Baumol and Maco Stewart [6, 1971] on this point.)

In general, the Carnegie school view of the firm may be said to be made up of two components, both of which are related to the problem of deviation from rational decision-making. These two elements are: (a) the ingredients that enter into the decision process and (b) the likelihood that various sub-objectives of the firm will be inconsistent. Frequently, the organizational sub-goals considered are (a) a production goal, (b) an inventory goal, (c) a sales goal, (d) a market share goal, and (e) a profit goal. Different members of the coalition, which is made up of those who run the firm, may espouse different goals, and bargaining takes place between them in order to solve this problem. However, except for the statement that organizational slack is likely to enter as a means of conflict resolution, the Carnegie school does not present a theory of how these various conflicting elements are resolved. Essentially two elements would be necessary in order to make the theory exist and "move," so to speak. One would be a theory of goal determination by the firm from components (sub-goals) taken from within the firm and the other a theory of conflict resolution of such sub-goals. Hence, it seems fair to say that while there are elements in the Carnegie approach that suggest the ingredients of specific micro-micro models, these elements are still incomplete, or the elements have not been selected and put together in such a way as to form a competing theory. As far as I could judge, the specific authors do not appear to claim much more for the Carnegie school approach.[2] Nevertheless, the Carnegie school approach has been extremely useful in providing the raw materials for simulation models based on assumptions that differ from the maximizing paradigm. Furthermore, as micro-micro theories are developed in greater depth, they are bound to use some of the existing Carnegie school ideas or ideas that overlap with them.

There is a large and growing literature that looks at various problems internal to organizations from an optimizing standpoint, and which are quite likely of interest to those concerned with micro-micro theory. We cannot review this literature (although such a review would be useful), but it may be helpful if we point to some examples. For the most part the authors do not ask what the behavior relations are but seek to provide optimal solutions to predetermined organizational problems. Thus James A. Mirrlees considers the optimal relations between reward administration and the monitoring of performance [35, 1976]. Joseph Stiglitz considers some related questions concerning internal incentives, lack of information about performance, piece rates, monitoring problems, *etc.* [48, 1975]. Roy Radner's and Jacob Marschak's work on the theory of teams is concerned with the information aspects of team work [41, 1972]. Radner and Michael Rothchild look at the allocation of effort from a similar standpoint [42,

---

[2] Baumol and Stewart [6, 1971, pp. 122–23] appear to agree with this assessment.

1975]. An essential aspect of these papers is to recognize an organizational problem and to find an optimum solution. It is of interest that the various papers do not deal with the question as to whether or not organizations seek, find, or use optimum solutions. A paper by Day and E. Herbert Tinney takes a somewhat different approach [14, 1968]. Their model has two departments, an engineering department whose objective is to minimize cost and a sales department with a profit maximizing objective. They show that a response to success and failure mechanism leads to dynamic behavioral trajectories consistent with the neoclassical model. It would be of interest to rework the exercise based on other behavioral assumptions for the two departments. On the whole this growing body of literature is of considerable importance, since it helps to fill out some of the theoretical possibilities about the nature of the "black box."

Another body of literature tangential but not unrelated to our problem seems to be inspired by Ronald Coase's famous article, which attempts to explain why firms should emerge [8, 1937]. Coase tried to answer the question in terms of the idea that firms save on transaction costs. Two especially noteworthy contributions are those of Armen Alchian and Harold Demsetz [2, 1972] and Oliver Williamson [51, 1975]. Alchian and Demsetz argue that as individual (cottage industry) production efforts become more complicated, monitoring the work or work results is required, and the firm provides internal monitoring activities more cheaply than do individuals working in isolation.

Williamson examined the problem much more extensively in *Markets and Hierarchies* [51, 1975]. There he analyzes the advantages of shifting transactions from the market to hierarchical structures. Williamson begins by analyzing the advantages of the most basic grouping of people within the firm, the peer group—advantages both to members of the group and to the overall efficiency of the firm. He then considers the advantages of vertical integration, which culminates in the multidivisional structure. While Williamson does acknowledge the possibility of conflict between workers and managers and is concerned with how the firm can elicit consummate rather than perfunctory cooperation from the workers, he does not discuss in any detail the inefficiencies of the peer group and/or hierarchical structures. While Williamson does not provide a micro-micro model, since that is not his problem, the book does contain innumerable useful ideas and insights for those interested in constructing micro-micro models.

IV. *General X-Efficiency Theory as a Micro-Micro Theory*

I have referred to my work that attempts to explain the phenomenon of X-efficiency as General X-Efficiency theory. In this section I want to consider how this work fits the concept of a micro-micro theory. Unlike conventional micro theory, it is unrealistic to assume that readers have read any of the expositions of X-efficiency theory. Hence this treatment will be disproportionately long. Of course there is no necessary connection between length and importance. At the same time I do not wish to use very much of this space to present another fairly full exposition. This part of the paper cannot be self-contained. What I wish to do instead is to indicate in a few pages (once again) a few basic elements of the theory to the extent necessary for purposes of this article. Readers interested in a fuller treatment can check the references.

We shall focus on five elements of X-efficiency theory: (1) selective rationality (degree of maximization deviation); (2) in-

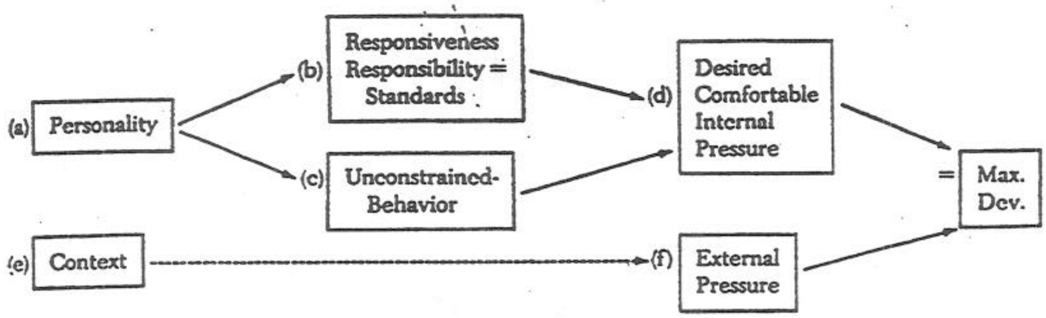

(a) Personality
(b) Responsiveness Responsibility = Standards
(c) Unconstrained-Behavior
(d) Desired Comfortable Internal Pressure
(e) Context
(f) External Pressure
= Max. Dev.

DIAGRAM 1

dividuals as basic decision units; (3) effort discretion; (4) inert areas; and (5) organizational entropy. In the neoclassical theory, maximization is the counterpart to selective rationality. The other elements are of a micro-micro nature and have no neoclassical counterparts.

Under selective rationality individuals select the extent to which they deviate from maximizing behavior. If two individuals *A* and *B* are faced with the same flow of opportunities and constraints (note that the constraints are attached to the opportunities) and *A attends* to all the opportunities and constraints, while *B* attends to only half of them, we can say that *B* deviated from maximizing behavior by 50 percent. The degree of maximization deviations depends on the elements indicated in the scheme shown on DIAGRAM 1. Personality is defined in terms of (b) a taste for responsiveness to opportunities and constraints within certain standards of behavior and (c) a simultaneous taste for "irresponsible" or unconstrained behavior. The standards of behavior include moral constraints reflected in attitudes toward trust, honesty, lying, altruism, group solidarity, sacrificing for group objectives, and so on, which play a role in behavior. The compromise between (b) and (c) that the personality makes leads to a most "comfortable" degree of internal pressure which, if everything were in the person's control, would determine the degree of maximization deviation. Maximization deviation is a measure of selective rationality. However, the economic context (e) in which the individual finds himself may impose a higher degree of external pressures than desired internal pressure, and hence the interaction between (d) and (f) determines the actual degree of maximization deviation. That is, pressures from peers (horizontal relations) and authorities (vertical relations) within the firm determine a degree of selective rationality different from what the person would choose in the absence of such pressures.

Now, we assume that firm membership contracts are incomplete.[3] The simplest case to consider is the one in which the individual is told what payment he is to receive but not how much effort he has to put forth. Hence effort discretion exists. Effort has four dimensions: activities *A*, pace of activities *P*, quality of acts *Q*, and time duration and sequence *T*. One value for each dimension yields an *APQT* bundle or effort point. Individuals will usually

---

[3] The fact that such contracts usually involve an indefinite future in which both reliability and flexibility are simultaneous requirements implies the necessity of effort discretion. See Ian R. MacNeil [28, 1974] and Victor Goldberg [18, 1978] on the significance of such contracts.

486

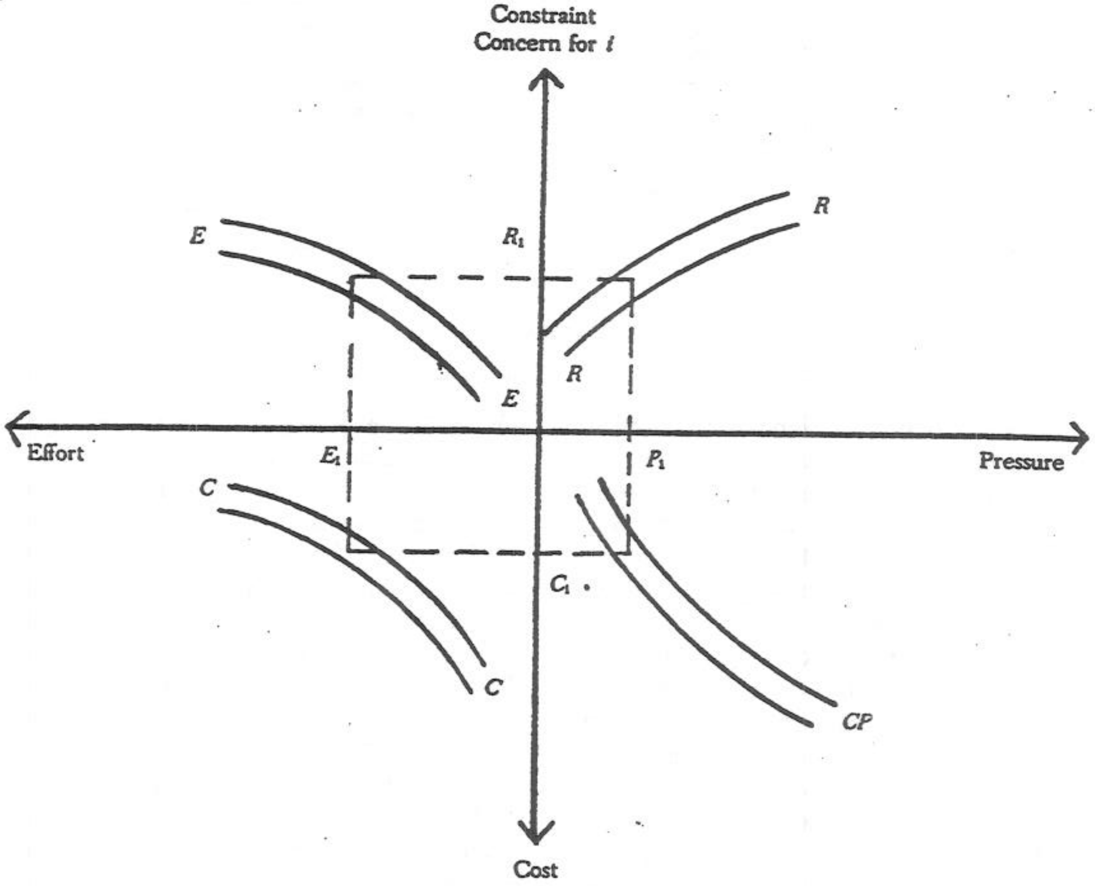


Figure 1

interpret their jobs (firm membership role) in terms of a subset of related effort points (to allow for variable demands on effort), which we will refer to as an effort position.

The idea of inert areas indicates the existence of effort points or of "positions" (whether effort positions or other types of option choices), so that if one is at one position associated with a certain utility one does not move to another position associated with a higher utility. The reason is that there is an "inertial cost" in moving from one position to another, and for the positions in the inert area, the inertial cost of moving is greater than or equal to the utility gain.

In diagramatical expositions inert areas will usually be reflected by "fat" curves. Now, it should be clear that the most important distinction between the X-efficiency theory and the neoclassical model of the firm is the fact that in X-efficiency theory the individual is the basic atomic unit. This by itself assures that the theory will be a micro-micro theory. If we add to this the assumption that effort discretion exists for every firm member, in part because employment contracts are incomplete, then clearly everyone in the firm helps to determine output, and hence each and every individual's choice helps to determine the cost of production.[4]

When an individual joins the firm, he is likely to be given bits of information that give him various clues that help him to guess the range of the area of effort discretion. In addition his perception of (1) the horizontal (or peer group), (2) the vertical (or hierarchical), and (3) the historical influences helps to guide him in his choice.[5] These constraining influences help to determine his choice. They are based on the likely reactions of others. That is, on the likelihood of approval or disapproval (1) on the basis of what is or is not done traditionally, (2) from peers, and (3) from those who are "authorities" or more specifically, those who are hierarchically above and below him.[6] These ideas are illustrated with the aid of *Figure* 1 above.

In quadrant I the abscissa P indicates the pressure that the firm imposes on the individual. The curve RR indicates how external pressure is translated into constraint concern by the individual. In quadrant II constraint concern becomes the independent variable, which in turn determines, through the function EE, the effort put forth. In quadrant III the curve CC translates effort into the cost of the production per unit contributed by the individual. In quadrant IV cost is the independent variable, and the curve CP shows how the firm reacts to the cost attributed to individual i by putting pressure on the individual. The greater the cost, the greater the pressure by the firm to reduce cost.

The essence of the relations in *Figure* 1 is shown in *Figure* 2. Constraint, concern, and effort are intermediate variables between pressure and cost. Hence, in *Figure* 2, in which the two variables are cost and pressure, the curve $C_i(P)$ indicates cost for the *representative* individual i as a function of pressure (suppressing R and E in *Figure* 1) and $P_i(C)$ as pressure imposed by the firm as a function of cost. The intersection between these two curves reflects an equilibrium between the representative individual i and the firm. If we start at a nonequilibrium point, we can see that there exists a cobweb pattern of reactions that leads back to the equilibrium position E in *Figure* 2. Other dynamic sequences are also possible. Space considerations do not permit a discussion of the relations between representative individuals, nonrepresentative individuals, and the firm.

Another idea related to effort discretion is organizational entropy. Entropy is used in the sense of a tendency towards disorganization. An individual chooses his effort position on the basis of what he expects the horizontal and vertical relations to be. However, if he discovers in the course of his activities that the horizontal and vertical constraints are much weaker than he anticipated, then he is likely to shift to those effort points within his effort position that he prefers, and which are less likely to be connected with the objectives of the firm. Hence we visualize this

---

[4] Mainline theory must assume implicitly that employment contracts are complete, or that supervision is precise and makes up for contractual incompleteness, if we are to have a conventional production function.

[5] We do not specify exact choice criteria nor do we assume a uniform choice procedure used by all individuals. Behind these ideas is the implicit assumption that different individuals will use different information search and different decision procedures. For each individual the search and decision procedures reflect the degree to which a person's behavior deviates from maximizing behavior. We do assume that once the activities reflecting these procedures fall into the inert area that this ends the search for options that can yield superior results. It would take us too far afield to go into details on these matters here.

[6] The authority-sanction-approval mechanism is an important part of the internal structure of the organization; *i.e.*, it is part of the insides of the "black box." The "voice" versus "exit" modes of behavior, analyzed by Hirschman [10, 1974] are part of this structural mechanism, both for imposing and for resisting pressure within the firm.

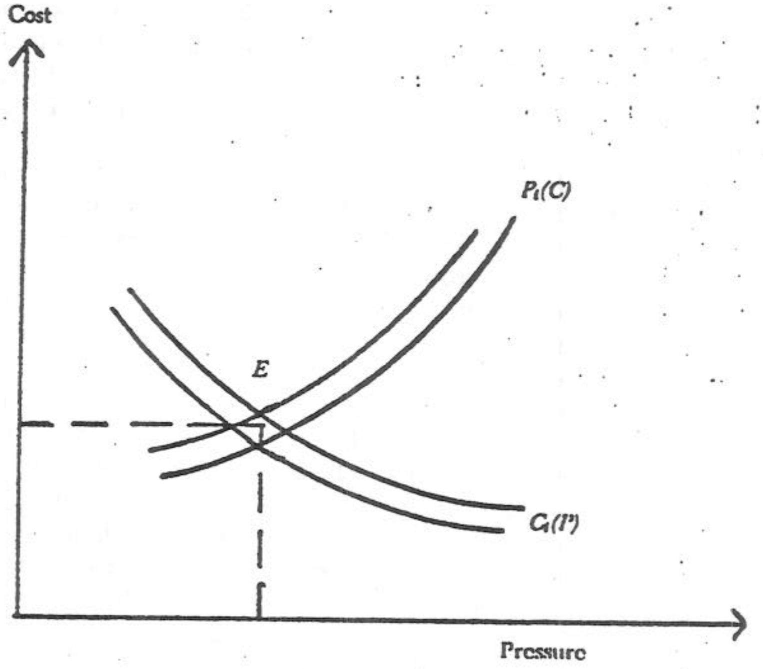


Figure 2

movement, when it manifests itself, as a process of organizational entropy. In *Figure* 2 it may be shown as a limited upward shift of the $P_i(C)$ curve. Thus we see in the figure that cost rises as entropy increases take place. However, an entropy increase may not manifest itself. We visualize the firm as an arena in which management struggles against the latent entropy force inherent in the organization and in effort discretion. If the management struggle is equal to the latent entropy then, of course, actual entropy increases will not be observed; *i.e.*, the curve $P_i(C)$ will not rise. But in any event we see the entropy force as a *latent* cost increasing process within the firm.

Figure 3 is similar to *Figure* 2 except that it aggregates the situation for all individuals. Aggregation is not simple. There is a distribution problem involved, which we will ignore in this brief exposition. For every level of aggregate pressure that the firm imposes, there is a distribution of pressures over individuals. Let us assume that pressure is the same for all individuals. We write p_{ij} to mean that for every individual i, the pressure is j. We can then obtain aggregate $P_j = \Sigma_i p_{ij} = n p_{ij}$ for n individuals. For each pressure level P_j we can now add up the costs and obtain the aggregate cost-pressure function C, and in a similar manner the aggregate pressure-cost function P.

The level of aggregate pressure P_j depends not only on cost but on influences outside the firm. Referring to the firm under analysis as firm A, if prices of competing firms are low compared to the cost level in a given firm A, then the members of the management team in A will feel more pressure to try to do something about costs than if prices in other firms were relatively high. We may visualize a

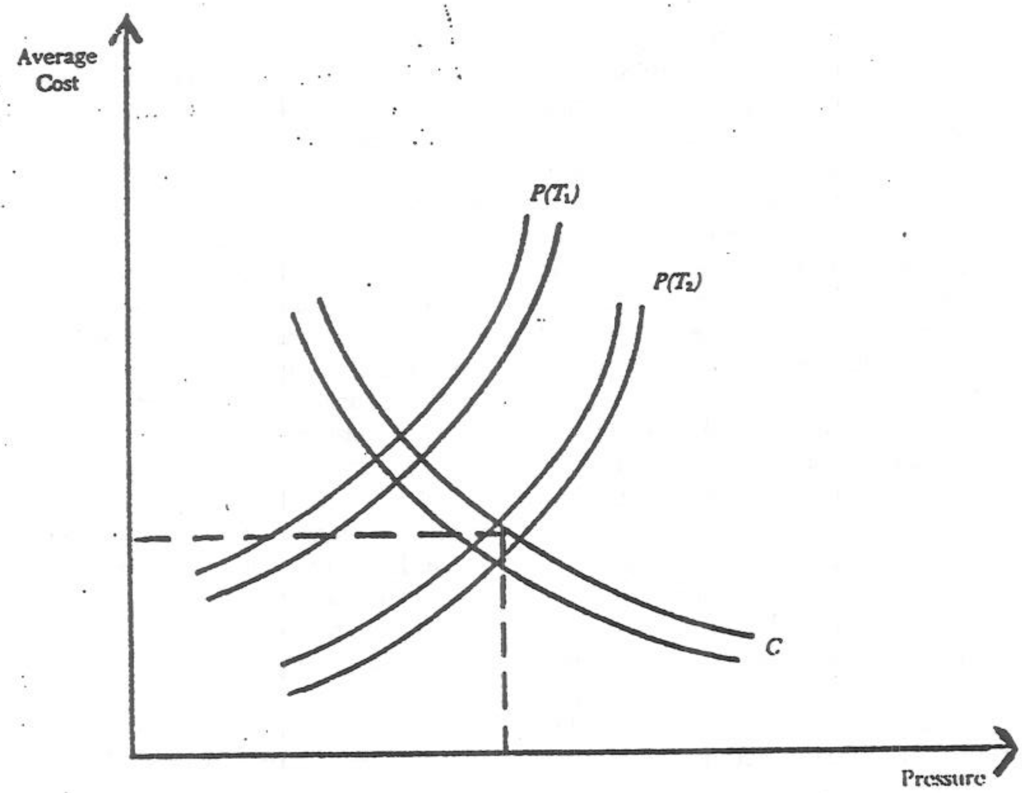


Figure 3

situation in which competing prices are low as a "tight" environment. Alternatively, if competing prices are high, it is a "loose" environment. Clearly the pressure that the management[7] imposes *inside* the firm will depend not only on cost but also on the level of environmental tightness. Thus for a tighter environment T_2 we show the P curve below what it would be for the looser environment T_1. As shown in *Figure* 3, equilibrium costs turn out to be lower under T_2 than under T_1. Of course, price is only one manifestation of environmental tightness—since we visualize environmental tightness as a more general concept than degree of competition. However, this is not the place to go into this matter.

Firms do not necessarily accept the degree of tightness—they can also attempt to determine it or contribute to its determination. We refer to activities whose aim is to reduce tightness as sheltering activities. In other words, the firm tries to find a shelter from competition for its high cost activities. First note that a substitute for reducing costs (fighting entropy) is to reduce tightness; *i.e.*, to change the environment. Thus pressure activities on firm members and sheltering activities are substitutes for each other. Examples of sheltering activities are the following: entering into price agreements, activities that help maintain price agreements, product differentiation, activities such as advertising, developing trademarks, engaging in market share agreements, entering into mergers, political activities to obtain price supports, tariffs or other restrictions on trade, and so on. The function of such sheltering activities is to reduce the impact of competition in the industry, or by possible entrants, in order to maintain or raise the existing price or prices charged by the firm in question. Thus, in general, the aim of sheltering activities is to increase the capacity to raise prices without excessive loss of sales. Of course, the firm in question need not initiate such activities. In part it may engage in such activities by acquiescing to or supporting the sheltering activities of others. Thus, in the price leadership situation following the price leader is a type of sheltering activity. Similarly, supporting political activities to increase a particular tariff is a sheltering activity that is likely to be approved by both management and trade unions that represent some or all of the workers within the firm. Thus sheltering activities are facilitated by various organizational networks, which permits the flow of support from inside the firm to industry-wide organizations, to political parties, and to governments.

---

[7] Every so often we use "management" as a linguistic shorthand to avoid getting into the problem of the distribution of activities among members of management teams. Of course, we have in mind that individuals impose these pressures on others.

From the point of view that we are taking, the opposite of sheltering activities are entrepreneurial acts—*i.e.,* those activities that set up new units, or expand existing units, in competition with existing firms. Thus anybody who sets up such units increases the response to an existing high price or to price increases. Hence, entrepreneurial activities have the effect of putting pressure on firms to reduce their prices. These activities are of course determined not by specific existing firms but by the entire economy. This is not the place to present a theory of entrepreneurship; however, consistent with X-efficiency theory we have to start with imperfect markets. (See Leibenstein [26, 1978, ch. 3].) Hence, our markets are characterized by *gaps* in input availabilities as well as obstacles that manifest themselves in less than equal access to inputs. Hence, in order to start a new firm, an entrepreneur must have the capacity to overcome the gaps and the obstacles in imperfect markets. He also has to be an *input completer* in the sense that less than the complete marshaling of inputs would not fulfill the entrepreneurial function. Obviously there are limits to the supply of gap fillers and input completers. Without going into detail as to all the elements that determine the supply of entrepreneurship, it seems likely that the higher the price, the greater the entrepreneurial supply. Also the lower the price, the greater the inducement to engage in sheltering activities.

For ease in exposition it is useful to consider simultaneously both sheltering and entrepreneurial activities. At any given price, sheltering activities would be aimed at raising or maintaining that price, while entrepreneurial activities would generally result in a lower price. However, an inadequate supply of entrepreneurial activities, *e.g.,* a level that would not replace all the firms going out of business during any period, would have the opposite effect. Now, both sheltering and entrepreneurial acts cost money, time, or both. Thinking of time in terms of the pecuniary value of its opportunity cost, we can use the monetary cost of entrepreneurial and sheltering acts as a measure of both. When we speak of both sheltering and entrepreneurial acts simultaneously, we shall refer to them as *environmental acts,* since they are essentially environment altering activities.

There are some intermediate steps in this analysis that will be left out. In a many-firm industry each firm contributes to sheltering or entrepreneurship or both, but no single firm can determine the environment. Hence what we illustrate in *Figure* 4 are the activities for all firms in

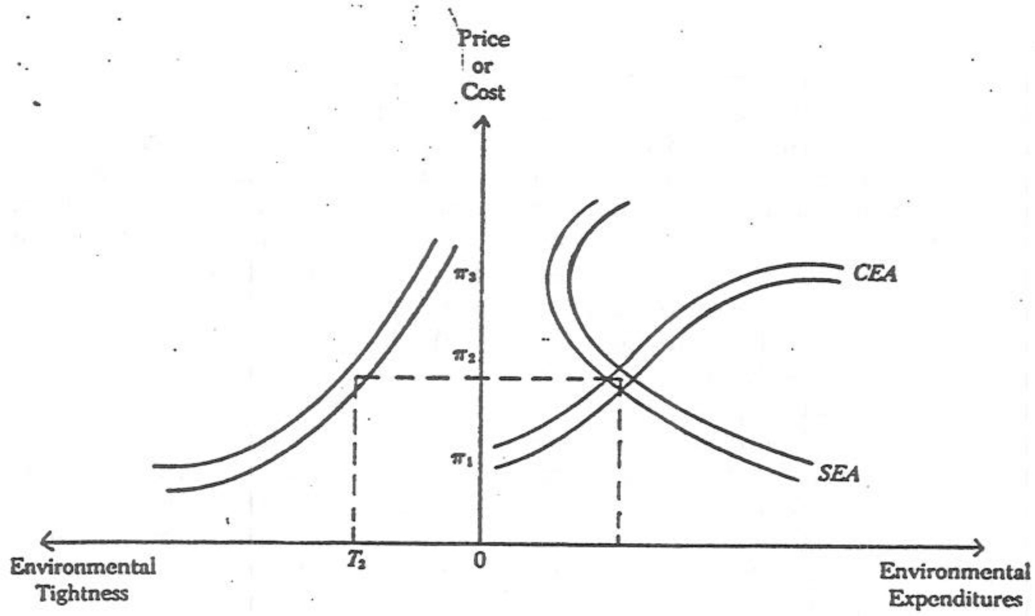


Figure 4

the industry, including that of latent entrepreneurs not currently in the industry.

It is also worth noting that with respect to sheltering activities a possible free-rider problem can arise. Each firm gains from the sheltering acts contributed by others. It is possible that there are cases for which the sheltering cost for each contributor is less than the gain to them if all firms do not contribute. A nonmaximizing model can help us out of the dilemma, since only non-profit-maximizers can be expected to act in such a way that they help to create a sheltered environment even if the gain of doing so is less than the cost when there is no assurance that others will contribute simultaneously. The reader will recall that the components of selective rationality include "moral" notions. Executives of firms may feel that their industry "ought" to be organized so as to weaken "unfair" competition.

In the right quadrant of *Figure* 4 we show the supply of environmental activities *SEA*, the consequences of environmental activities *CEA*, and the interaction between these two functions, which determines the equilibrium degree of tightness in the market. The abscissa measures environmental acts in terms of expenditures on such activities, while the ordinate measures the expected long run average price. The shape of the *SEA* curve reflects the expectation that at high prices entrepreneurial acts dominate, while at low prices sheltering acts dominate; while at some level of intermediate prices there is neither great enthusiasm for entrepreneurship nor much-felt need for sheltering. Now the curves *CEA* reflect the consequences of environmental acts.[8] We

---

[8] Entrepreneurial *(E)* and sheltering *(S)* acts are each assumed to be separate functions of price. Each point on the *SEA* curve reflects the sum of the entrepreneurial and sheltering acts at that price. Thus each point on the ordinate reflects a determinate mix of *E* and *S* acts. Points on *CEA* reflect the consequences of that mix as reflected in the *SEA* curve. A more detailed theory would start with a model of the determination of the specific contribution of each firm to the *SEA* curve. Space limitations prohibit a consideration of this problem here.

442 consider the *CEA* function only between prices π_1 and π_3. We assume that this is the segment of sustainable prices. Above π_3 the consequences of the mixture of entrepreneurial and sheltering acts are such that the price cannot be sustained. Hence no values on the *CEA* curve are shown above π_3. The shape of *CEA* presumes that the consequences of sheltering activities dominate as prices rise from π_1 to π_3. *CEA* would be downward sloping if the opposite were the case. The point of intersection at π_2 determines the equilibrium level of environmental activities, and its related degree of tightness is shown in the left quadrant. This analysis could be developed in much greater detail, but for present purposes it is sufficient to simply indicate that there are incentive elements, which generate activities that determine environmental tightness, which is then transmitted to the firm in question. This determines the pressure-cost relationship, which in turn gives the cost-price relationship, which in its turn determines the cost area of the commodity.

For present purposes we focus attention on *Figures* 1, 2, and 3, which are concerned with the micro-micro aspect of the theory. We note that the choice of effort positions by individuals essentially determines the cost attributed to each individual. Of course, the degree of organizational entropy and the extent to which managers and their effort activities are induced to fight entropy are also part of the firm picture. In addition the reader should recall that all curves are drawn in the four figures as "fat curves" in order to reflect the existence of inert areas. The significance of this is considered later.

Our theory also allows for the special case in which entrepreneurship is so vigorous that it induces a very high degree of tightness, which in turn forces a degree of pressure so that firms either minimize costs or do not survive. However, we must note that this is a very special case, one not likely to exist in most industries. Thus the micro-micro theory enables us to distinguish the environmental micro-micro conditions for perfect competition as against what is probably the normal situation of some form of imperfect competition, and hence the case in which costs are not minimized.

We have argued in an earlier section that one of the problems a micro-micro theory is likely to face is the resolution of possibly conflicting choices by firm members. We now consider very briefly to what extent X-efficiency theory handles this problem. A primary resolution mechanism lies in the existence of inert areas. Without inert areas the choices of *A* and *B* could easily conflict if they impinge on each other; but with inert areas many choices made by *A* that affect *B* lead to no response by *B*, and the same is true of *B*'s choices that affect *A*.[9] The larger the inert areas, the larger the nonconflict area, or mutual non-response area. If the inert areas do not resolve potential conflicts, at least two other mechanisms exist that help. These are horizontal or peer group pressures and vertical pressures, especially that of superiors on subordinates. Finally the system of authority and the persuasive sanctions available to authorities come into play. These mechanisms are usually sufficient for the day-to-day operations of the firm. Of course, some areas of conflict remain, and some extraordinary mechanisms may have to be used, *e.g.*, external mediation or arbitration in the settlement of strikes.

This general set of ideas does lead to some different conclusions than the neoclassical model. Among these are: (1) the cost of a commodity is not independent of the price of the commodity; (2) except in the extreme circumstances firms do not minimize costs; (3) cost of production has a tendency to rise toward the price level; and (4) there is no production function independent of the environment of the firm and the history of the firm.

---

[9] Mutually consistent inert areas would also allow for the persistence of repressed conflict in the organization.

V. *To Maximize or Not to Maximize*

In considering alternatives to conventional micro theory, one of the questions that arises is how to handle the usual maximization assumption. At present the maximization postulate has an unusually strong hold on the mind set of economists. To examine why this is the case would take us too far afield into the sociology of the profession. Suffice it to say that in my view the belief in favor of maximization does not depend on strong evidence that people are in fact maximizers. In this section we will (a) consider some arguments usually put forward in favor of the maximization postulate and (b) suggest some reasons why it is reasonable to relax the maximization assumption in developing micro-micro theories.

It should be clear on the basis of what has been argued heretofore that micro-micro theory does not depend on dropping the maximization assumption—but it helps. In part, this point will be argued in this section. However, the essence of micro-micro theory is the behavioral unit; *e.g.*, shifting to the individual rather than to the firm or to the household. Clearly a theory of the firm can start with the assumption that each individual member maximizes something or other, the simplest assumption being the maximization of utility. There are problems with this idea, but let's ignore these while we examine some possibilities and implications.

We may ask (1) whether individual utility maximization will or will not be consistent with firm maximization, and (2) what it is that firms maximize. Consider the case of a two-person firm. One possibility is for the two to attempt to achieve a joint maximum. The nature and limitations of this approach are already exemplified in the Edgeworth box diagram, in the large literature on bargaining, and in von Neumann and Morgenstern's *Theory of Games* [38, 1944]. Such situations are usually characterized by simultaneous possibilities for *both* cooperation and conflict. The so-called contract curve in the box diagram indicates that for every choice off the curve there is a superior position on it. But once on the curve the cooperative aspects are exhausted, and movement along the curve implies that a gain for one will be counterbalanced by an equal loss to the other. Thus, even if it were in the interest of firm members to maximize the size of the pie, they would still face a pie-dividing problem. Furthermore, the situation need not be a zero-sum game. Tactics that determine pie division can affect the size of the pie. It is this latter possibility that is especially significant. That is, one of the questions we have to raise is whether, if we want an individual maximization theory, that theory should be allowed to reflect the possibility that the struggle over pie division can affect size.[10]

An argument could be developed to the effect that individual utility maximization would destabilize intrafirm behavior. Suppose that every member of a firm is a utility maximizer. How would such people behave? By definition they would respond to every "data" change that would affect them. Since such data changes are numerous and frequent in most firms and markets, such individuals would be constantly interested in changing their jobs inside or outside the firm. At the same time they would be interested constantly in renegotiating their contracts. The transaction costs created by the constant search for even a slight advantage may be low for each individual, but quite high in the aggregate for the firm as a whole. One could easily develop a model the result of which would be that firms would prefer to hire people who would act *as if* they had inert areas of a reasonable size. Better still, the firm would prefer sincere non-maximizers.

---

[10] A somewhat related way of looking at this problem is to note that intrafirm behavior involves free-rider elements and that it can be characterized as a prisoner's dilemma type of problem. But there is frequently a resolution of such problems if we take "ethical" or superego considerations into account. Note that these elements are contained in the non-maximizing selective rationality approach. Each may prefer the other person to do the work, but if each has a strong sense of a "fair" day's work, despite the fact that it is against his best interest to do so, then the free-rider prisoner dilemma aspects of the problem can be overcome. The reader will also note that in our discussion of sheltering similar free-rider prisoner's dilemma situations arise, and similar possible resolutions may exist.

The main argument against the maximization postulate is an empirical one—namely, people frequently do not maximize. Of course, this standpoint argues that while postulates simplify reality, we are not free to choose counterfactual postulates. Hence, from this point of view a superior postulate would be one under which maximizing behavior is a special case, but non-maximization is accommodated for as a frequent mode of behavior.[11]

We now turn to some of the arguments usually brought forth in favor of the maximization hypothesis. One point must be made clear before we start. We shall not argue against maximization as a description of behavior under the circumstances under which such a description is accurate. In other words, the view to be taken is that maximization as a description of behavior is an empirical question. It is either universally true or not. I believe there is evidence that it is not universally true. On the whole it is rare to find anyone arguing that maximization is a universal description of behavior despite the ubiquity of the assumption. It is an assumption that is most frequently stated rather than supported.

A position frequently taken is that maximization is invariably true if only we take into account (a) the right considerations and/or (b) enough considerations. Thus if on the basis of some consideration it appears that maximization has not taken place, then it is only because an additional consideration has not been taken into account. Thus if an economic agent has not appeared to have maximized profits, it is only because leisure has not been considered. If it had been, it would be reasonable to presume that the agent had maximized the utility of a function made up of both profits and leisure. Similarly, if a two-argument function does not appear to have been maximized, then a three-argument function would solve the problem. And so the argument goes. This type of thinking has an extremely strong hold on the economics profession. In one form or another I have heard it put forth in a number of seminars. Yet this argument has serious flaws. We now turn to examine the form of this argument, and the appropriate response to it, although we cannot examine all variants of the argument.

(1) The first point to be considered may appear to be linguistic, although in this case it is also logical. Words have to be used in such a way that they do not contradict their essential meaning. Thus, in using the word "maximization," we must of necessity assume that non-maximization is meaningful in the same context. Otherwise the word maximization is robbed of its meaning. And obviously we are not using the word maximization appropriately if we rob it of its meaning. If we say that there are three options, *A*, *B*, and *C*, valued at 3, 7, and 10, then the person who maximizes chooses *C* valued at 10. But by the same token we must imply the *possi-*

---

[11] One may argue in favor of a maximization postulate as a convenient research strategy. This may be best for some problems and not for others. It seems to me one cannot say anything on this question *a priori*.

bility of non-maximization; that is, someone could choose options *A* or *B*.

Now suppose someone chooses option *B*. How shall we judge such a choice? One possibility is to say that this represents non-maximizing behavior. But what of the argument that this really reflects maximizing, but that other considerations are involved so that the valuation of *B* is really greater than *C*, and furthermore, that similar arguments apply in *all* cases. Now this approach goes counter to the appropriate use of words. This approach *essentially* denies the possibility of non-maximization, since every choice that is made is interpreted as maximization; it robs the assertion of "maximizing" of meaning. It says no more than people behave as they do. Giving the name "maximization" to what people do does not add information. In fact it is most likely to hide information and to deter questions about what is actually happening.

An approach sometimes used explicitly is the tautological interpretation of utility maximization. This is similar to the approach considered in the previous paragraph except that it has the virtue of being explicit rather than implicit. From this standpoint not only is everyone assumed to maximize all the time, but non-maximizing behavior is *assumed* to be out of the question. Essentially, this says that whatever people do, they are somehow maximizing utility, without specifying the components they add or subtract from the individual's utility. Since all behavior is subsumed under this view, it essentially says nothing more than people do what they do. There is an extraneous assertion that is added on to the notion that people do what they do, which suggests that when they do what they do they are also maximizing utility. Clearly the concept that people do what they do is non-falsifiable and unrelated to the proper use of the word maximization.

(2) A nontautological approach must lead to the conclusion that whether or not maximization takes place is an empirical question. This is not to suggest that with present methods we can always determine whether or not maximization is in fact taking place.[12] Rather, it simply says that the appropriate use of the term of maximization or non-maximization involves the idea that in principle one could check it on the basis of empirical data. Obviously not all empirical questions can be handled with current techniques.[13] This is true at some time or other in all sciences.

Herbert A. Simon in his 1978 Ely lecture argues that where scarce decision resources (*e.g.*, computational, search, or attention means) are involved, optimizing decision procedures may not be rational. Thus, he argues that an ". . . important direction of research in computational complexity lies in showing how the complexity of problems might be decreased by weakening the requirements for solution—by requiring solutions only to approximate the optimum, or by replacing an optimality criterion by a satisficing criterion" [47, 1978, p. 12]. This yields an additional reason for not using the maximizing postulate.

(3) If maximization is to include all types of behavior and all types of tastes, we have to raise the question of how we should handle a taste for nonmaximizing behavior. In that case we would come up against assertions of the following types: That a person is maximizing when he behaves in accordance with his taste for non-maxi-

---

[12] Of course in some cases we can tell. In his Ely lecture [47, 1978], Herbert A. Simon cites the work of Howard Kunreuther *et al.* to the affect that decisions to buy flood insurance depended more on personal experience with floods than with a knowledge of the rates, objective risks, or even subjective beliefs about these facts.

[13] A desirable convention that would avoid some confusion would be for writers to indicate explicitly whether they are using maximization in a tautological or nontautological sense.

mizing. Obviously there is a contradiction in the notion that a person is maximizing when he is not maximizing.

(4) In the discussion in the previous section, we have suggested a nonmaximization model in which statements were made about the psychological attributes of an individual and about inertial costs. It has sometimes been argued that when such costs are taken into account that what appears to be a nonmaximization approach turns out to be maximization. Such an approach involves a very significant distinction: The distinction between statements about the personality of an individual and statements about his behavior. To say that someone's inertial cost is clearly zero is a statement about his personality. It indicates how he feels about making a certain type of choice. It does not discuss how an individual behaves when faced with alternatives. The inertial cost idea explains *why* a person behaves in a certain way. It does not indicate *how* he behaves. The statement that a person is a non-maximizer is an assertion about *how* he behaves. If these two elements are confused, the "why" and the "how," then we end up with a tautological interpretation of maximization, an interpretation that robs the concept of maximization of its meaning.

(5) A very important element in any conceptual scheme is whether the scheme allows us to ask interesting questions. The difficulty with assuming maximization is that within this assumption we cannot raise the question of what happens when behavior is of a non-maximizing sort and, even more important, we cannot raise the question of what happens during the interaction between those individuals who maximize and those who don't. Does a maximizer have an advantage over a non-maximizer? Can the careful calculator victimize the incomplete calculator? Clearly there are a host of interesting questions involving such interactions that should be investigated. It may turn out that part of the important element of micro-micro theory depends on interacting behavior between individuals who are calculating to very different degrees.

(6) We have argued in previous sections that conflict resolution is likely to be an important aspect of micro-micro behavior. In a sense, maximizing behavior by interdependent individuals is the opposite of conflict resolution; whereas some forms of non-maximizing behavior allow for considerable latitude of accommodation between competing objectives.

Dropping the maximizing postulate is likely to raise a host of interesting questions; one of the most important of which is the determination of the set of decision processes actually used within a group. This question is brushed under the rug, so to speak, by assuming maximization. It may turn out that the mix of decision processes used within a firm, or a group of firms, has important economic implications. We do not know the answer, but it would seem exciting to open up the question.

VI. *Summary and Conclusions*

In a number of the natural sciences, at various times, there has been a special sense about some areas of study—the sense of a *research frontier*. I am told that in physics, microbiology, and in some branches of chemistry, there is a clear sense of such a frontier. Investigators in these fields frequently know and agree where fruitful work can take place, and they usually have a lively sense of where in fact it is taking place. James Watson's book, *The Double Helix* [49, 1968], dramatizes those aspects in which there is clear-cut convergence of a number of investigators in a number of disciplines, almost inexorably toward the discovery of the structure of DNA and the relationship of the structure to its capacity to replicate itself.

Economics as a science has never had a clear sense of a research frontier. Investigators work in a variety of subdisciplines without much convergence. Economics as a discipline has changed gradually, but there is still considerable controversy as to whether the new stuff is any truer than the old. As a profession we have become more mathematical, but that does not mean we know any more about the world of economic phenomena. The closest thing to the opening up of a research frontier occurred in the 1930's in the wake of the so-called Keynesian revolution. Probably a better name would have been the macroeconomic revolution. While there is still much to know, the work of an army of investigators resulted in changing the face of economics—basically, the formalization of *macro*economics. The current spate of multiequation forecasting models, which depend on computers to derive results quickly, may be looked upon as the end product of this type of work. But outside of the macroeconomic breakthrough of the thirties, there is very little that has occurred in the twentieth century that we can point to as a research frontier.

The point of the above discussion on the idea of a research frontier is to suggest that micro-micro theory has the potential for becoming such a frontier. To start with, a great deal of accumulated information already exists on the major entity to be studied—the individual. Biologists, psychologists, and sociologists already know a great deal about individual behavior in a variety of contexts, including economic contexts. The findings in the field of industrial psychology may be useful to our understanding of how individuals function in firms. But much more important than the actual facts about individual behavior are the various *techniques* developed in these disciplines to study individual behavior in organizational contexts. Every discipline develops its own special questions, and the questions most pertinent to industrial psychologists or sociologists may not be the most pertinent ones for economists. While we frequently cannot use the answers developed in these fields, we can use their methods of inquiry, albeit refined to our own needs and standards.

Microeconomics has avoided the study of individuals, but that is no reason why we should continue to do so. After all, once we change the central questions to the ones pertinent to micro-micro theory, we can investigate the empirical elements. At the very least we can observe individual behavior in economic contexts. Also, we can submit individuals to questionnaire surveys, as well as to try to observe them in experimental settings. The main point is that individuals are the most fundamental and the easiest units to observe. Once economists really get into micro-microeconomics, they can gather their own data, the data exactly relevant to their concerns and questions; test their hypotheses; and obtain results.

The basic point is that individuals within organizational contexts are likely to be easier to study than individuals generally. Furthermore, individuals are easier to study than multiperson firms as units. This argument is difficult to develop rigorously. No such attempt will be made. Rather, we hope to indicate some of the components of the argument in the hope of suggesting that the general argument is a reasonable one. It would seem almost self-evident that it would be easier to study individuals in limited sets of contexts rather than in all contexts. In some sense this implies that by choosing our contexts very carefully we may be able to get close to laboratory experiments. The ideal would be to find organizational contexts that are as close to being identical as possible and to see in what way individuals behave differently, given different values of the independent variables in such contexts. Obviously there are likely to be different degrees to which it may be possible to find contexts that are similar, but in which only one of the few independent variable values differs. In general it would seem that by reducing the variety of behavioral contexts, it should be very much easier to study individual behavior.

The argument is somewhat more difficult when we compare the study of individuals versus the study of firms. One would think it would be easier to find units of firms that have greater similarity to each other than would firms as entities. For example, it might be easier to find similarity in plants than in firms, or in some sub-units of plants, than in firms. Hence, through the careful selection of some sub-units, it would be easier to study individuals than through the study of gross aggregate entities as units. Furthermore, once we accept the idea that we should look at what goes on inside the firm, it is almost self-evident that there are some researchable areas inside the firm. Of course, these two approaches, the study of the individual and the firm as units, are not by any means mutually exclusive. We would normally expect that information gained from each approach separately would turn out to be complementary.

Near the boundary of micro-micro theory there exists a number of models, most of which fall into the category of multivariable objective functions. The firm, or some persons or group within the firm, is presumed to maximize a utility function with many arguments. Typical arguments are growth, profits, managerial perquisites, leisure, and so on. We argued that while these models are of considerable interest in that they relax some of the conventional assumptions of micro theory, they may be viewed only as embryonic micro-micro theories. Although these models do separate owners from managers and are concerned, to some degree, with the relations between these two groups, they do not handle the problems of potential conflict within multiperson managerial groups. Other studies, such as the theory of teams, are also just within the micro-micro boundary. The team may be a portion of a firm rather than the entire firm, but such studies are usually normative or prescriptive, rather than behavioral.

The work of the Carnegie school can be interpreted as being within the micro-micro framework, but the Carnegie framework is at present too incomplete a system to permit determinate conclusions. The simulation models inspired by or based on Carnegie school ideas are usually single case models. Other works that fall into this general category, although not directly based on the Carnegie school or considered here (*e.g.*, Eliasson, Nelson, and Winter, *etc.*) represent deviations from neoclassical theory, which yield results by using different decision criteria. However, these works are not primarily concerned with disassembling the black box of standard theory. They are, nevertheless, within the boundaries of micro-micro theory.

I have argued that my own work, which I have referred to as General X-efficiency theory, is a micro-micro theory. Since the individual is the irreducible analytical unit, models developed on this basis fit under the micro-micro rubric. At the present stage of development its focus is primarily on the factors responsible for cost determination. The basic decision-makers are *individuals* who, through their effort-discretion choices, determine cost of production.

In our discussion of various alternatives to neoclassical theory, the problem of giving meaning to "group utility functions" turned out to be troublesome. Intra-firm discretion of various sorts is a related or similar characteristic of many models. In both cases there are problems of resolving conflicts. We cannot give a general answer, but one form of resolution is through the notion of inert areas. If one's position falls into an inert area, conflict is no longer pursued.

Micro-micro theories, as they now stand, are far from being as well developed as neoclassical theory—especially the perfect competition case. In part this is one reason why it may be a candidate to become a fruitful research area. Different scholars will approach this area from different standpoints. In my view progress is most likely to be made if we drop at least some of the assumptions of conventional theory. Obviously the most important assumption that has to be dropped is the one under which firm behavior is visualized as if it is the behavior of a single individual.

An area where, in my experience, a good deal of controversy is likely to arise is the relaxation of the maximization hypothesis. In *Section* V an argument was presented in favor of relaxing this particular postulate. Nevertheless, it seems to me that at the present stage this is not an either/or question. It may be useful to work out models either on the basis of utility maximization or on the basis of nonmaximizing behavior. My own preference is toward models in which maximization represents a special case within a larger range of possibilities. In part, this reflects the notion that it would be useful to experiment with interrelations within firms in which some individuals are more calculating than others.

One of the difficulties in working in this area is that at the present time many economists appear to have strong feelings about the maximization hypothesis. Although strong feelings are understandable, given a century-long tradition within the profession of using maximizing-optimizing models and given the handiness of the mathematical tools available for maximizing models, nevertheless, they are not sufficient reasons to ignore the possibility of nonmaximizing models. Such models should be judged by their results and not by the criterion as to whether or not the model uses a specific hypothesis such as the maximization postulate.

Given that most sciences have moved from the study of less to more micro aspects, it would seem reasonable to cultivate this area in economics—especially to study intra-firm units and intra-firm individual behavior. But in addition to following a general scientific trend, there are intrinsic reasons to intensify our work in the micro-micro area. A body of work already exists, which suggests ways of studying these problems in greater depth. We should take advantage of the fact that we can observe intrafirm behavior and study such behavior intensively. A great deal of literature already exists in allied fields, such as industrial psychology, some of whose results can be incorporated into intra-firm models. And last but not least, it would seem natural at this stage to pursue work on micro-micro elements in order to check the consistency of the implications of what we know and what can be learned about intra-firm behavior with the postulates *and* implications of mainstream micro theory.

REFERENCES

1. ALLISON, GRAHAM T. "Conceptual Models and the Cuban Missile Crisis,". *Amer. Polit. Sci. Rev.*, Sept. 1969, *63*(3), pp. 689–718
2. ALCHIAN, ARMEN A. AND DEMSETZ, HAROLD. "Production, Information Costs and Economic Organization," *Amer. Econ. Rev.*, Dec. 1972, *62*(5), pp. 777–95.
3. ARGYRIS, CHRIS. "Management Information Systems: The Challenge to Rationality and Emotionality," *Manage. Sci.*, Feb. 1971, *17*(6), pp. B275–92.
4. ARROW, KENNETH J. AND HAHN, FRANK H. *General competitive analysis.* San Francisco: Holden-Day, 1971.

5. BAUMOL, WILLIAM J. "The Revenue Maximization Hypothesis," in *Business behavior, value and growth.* New York: Macmillan, 1959, pp. 45–53.
6. _____ AND STEWART, MACO. "On the Behavioral Theory of the Firm," in ROBIN MARRIS AND ADRIAN WOOD, eds. [32, 1971], pp. 118–43.
7. CAIRNCROSS, ALEC. "The Optimum Firm Reconsidered," *Econ. J.*, Supplement, March 1972, *82*(325), pp. 312–20.
8. COASE, RONALD H. "The Nature of the Firm," *Economica*, Nov. 1937, (4), pp. 386–405.
9. COHEN, KALMAN J. AND CYERT, RICHARD M. *Theory of the firm: Resource allocation in a market economy.* Second edition. Englewood Cliffs, N.J.: Prentice-Hall, [1965] 1975.
10. CYERT, RICHARD M. AND HEDRICK, CHARLES L. "Theory of the Firm: Past, Present, and Future; An Interpretation," *J. Econ. Lit.*, June 1972, *10*(2), pp. 398–412.
11. _____ AND MARCH, JAMES G. *A behaviorial theory of the firm.* Englewood Cliffs, N.J., Prentice-Hall, 1963.
12. DAY, RICHARD. "Review of Cyert and March, *A Behavioral Theory of the Firm,*" *Econometrica,* July 1964, *32*(3), pp. 461–65.
13. _____ *Behavioral economics.* Madison: University of Wisconsin, Social Systems Research Institute, 1975.
14. _____ AND TINNEY, E. HERBERT. "How to Cooperate in Business without Really Trying: A Learning Model of Decentralized Decision Making," *J. Polit. Econ.*, July–August 1968, Pt. I, *76*, pp. 583–600.
15. DEBREU, GERARD. *Theory of value.* New York: Wiley, 1959.
16. EDWARDS, RICHARD C. "Individual Traits and Organizational Incentives: What Makes a 'Good' Worker?" *J. Human Res.*, Winter 1976, *11*(1), pp. 51–68.
17. ELIASSON, GUNNAR. *Business economic planning.* New York: Wiley; Stockholm: Swedish Industrial, 1976.
18. GOLDBERG, VICTOR P. "Protecting the Right to Be Served By Public Utilities," in *Research in Law and Economics.* Vol. 1. Forthcoming, 1978.
19. HIRSCHMAN, ALBERT O. "Exit, voice, and loyalty: Further reflections and a Survey of Recent Contributions," *Social Science Information,* Feb. 1974, *13*(1), pp. 7–15.
20. JENNER, R. A. "An Information Version of Pure Competition," *Econ. J.*, Dec. 1966, *76*, pp. 786–805.
21. KNIGHT, FRANK H. *Risk, uncertainty and profit.* Boston: Houghton Mifflin, 1921.
22. LEIBENSTEIN, HARVEY. "Organizational or Frictional Equilibria, X-Efficiency and the Rate of Innovation," *Quart. J. Econ.*, Nov. 1969, *83*(4), pp. 600–23.
23. _____ "Notes on X-Efficiency and Technical Progress," in *Micro aspects of development.* Edited by ELIEZER B. AYAL. New York: Praeger, 1973, pp. 18–38.
24. _____ "Aspects of the X-Efficiency Theory of the Firm," *Bell J. Econ.*, Autumn 1975, *6*(2), pp. 580–606.
25. _____ *Beyond economic man.* Cambridge, Mass.: Harvard University Press, 1976.
26. _____ *General X-efficiency theory and economic development.* New York and London: Oxford University Press, 1978.
27. LINDBECK, ASSAR. "The Efficiency of Competition and Planning," in *Planning and market relations.* Edited by MICHAEL KASER AND RICHARD PORTES. London: Macmillan, 1971, pp. 83–107.

28. Macneil, Ian R. "The Many Futures of Contracts," *Southern California Law Rev.*, May 1974, *47*(3), pp. 691–816.

29. Margolis, Julius. "Analysis of the Firm: Rationalism, Conventionalism and Behaviorism," *J. Bus.*, July 1958, pp. 187–99.

30. Marris, Robin. "A Model of the Managerial Enterprise," *Quart. J. Econ.*, May 1963, *77*, pp. 185–209.

31. ——— "An Introduction to Theories of Corporate Growth," in Marris and Wood, eds. [30, 1971], pp. 1–36.

32. ——— and Wood, Adrian, eds. *The corporate economy: Growth competition and innovative potential.* Cambridge, Mass.: Harvard University Press, 1971.

33. Marshall, Alfred. *Industry and trade.* London: Macmillan, 1923.

34. Meyer, John R. "An Experiment in the Measurement of Business Motivation," *Rev. Econ. Statist.*, August 1967, *49*, pp. 304–18.

35. Mirrlees, James A. "Optimal Structure of Incentives and Authority within an Organization," *Bell J. Econ.*, Spring 1976, *7*(1), pp. 105–31.

36. Nelson, Richard R. and Winter, Sidney G. "Toward an Evolutionary Theory of Economic Capabilities," *Amer. Econ. Rev.*, May 1973, *63*(2), pp. 440–49.

37. ——— and Winter, Sidney G. "Neoclassical *vs.* Evolutionary Theories of Economic Growth—Critique and Prospectus," *Econ. J.*, Dec. 1974, *84*(336), pp. 886–905.

38. von Neumann, John and Morgenstern, Oskar. *Theory of games and economic behavior.* Princeton, N.J.: Princeton University Press, 1944.

39. Olsen, E. Odgers, Jr. "The Effort Level, Work Time, and Profit Maximization," *Southern Econ. J.*, April 1976, *42*(4), pp. 644–52.

40. Penrose, Edith T. *The theory of the growth of the firm.* Oxford: Oxford University Press; New York: Wiley, 1959.

41. Radner, Roy and Marschak, Jacob. *The economic theory of teams.* New Haven: Yale University Press, 1972.

42. ——— and Rothschild, Michael. "On the Allocation of Effort," *J. Econ. Theory*, June 1975, *10*(3), pp. 358–76.

43. Rasmusen, Hans J. "Multilevel Planning with Conflicting Objectives," *Swedish J. Econ.*, June 1974, *76*(2), pp. 155–70.

44. Scitovsky, Tibor. "A Note of Profit Maximisation and Its Implications," *Rev. Econ. Stud.*, 1943, *11*(1), pp. 57–60.

45. Simon, Herbert A. "A Comparison of Organisation Theories," *Rev. Econ. Stud.*, 1952, *20*(1), pp. 40–48.

46. ——— "Theories of Decision-Making in Economics and Behavioral Science," *Amer. Econ. Rev.*, June 1959, *49*, pp. 253–83.

47. Simon, Herbert A. "Rationality as Process and as Product of Thought," *Amer. Econ. Rev.*, May 1978, *68*(2), pp. 1–16.

48. Stiglitz, Joseph. "Incentives, Risk and Information Notes Towards a Theory of Hierarchy," *Bell J. Econ.*, Autumn 1975, *6*(2), pp. 552–79.

49. Watson, James D. *The double helix: A personal account of the discovery of the structure of DNA.* New York: Atheneum, 1968.

50. Williamson, Oliver E. "A Model of Rational Managerial Behavior," in Cyert and March [11, 1963], pp. 237–52.

51. ——— *Markets and hierarchies: Analysis and antitrust implications: A study in the economics of internal organization.* New York: Free Press, 1975.

52. ______. "The Economics of Internal Organization: Exit and Voice in Relation to Markets and Hierarchies," *Amer. Econ. Rev.*, May 1976, *66*(2), pp. 369–77.

53. WINTER, SIDNEY G. "Economic 'Natural Selection' and the Theory of the Firm," *Yale Econ. Essays*, 1964, *4*(1), pp. 225–72.

54. ______. "Satisficing, Selection, and the Innovating Remnant," *Quart. J. Econ.*, May 1971, *85*(2), pp. 237–61.

4-6

4

Handbook of Business Strategy, 2nd Ed.
Harold E. Glass, ed.
Warren, Gorham + Lamont, 1991

WHARTON REPROGRAPHICS

15

4

The Role of Science in Business Strategy

SIDNEY SCHOEFFLER
Founding Director, The Strategic Planning Institute

Why Examine the Issue? 2

What Is "Science"? 3
Observation and Evidence 3
Analytical Rigor 3
Significance 3
Probability 4
Hypothesis and Test 4
Representative Cross Section 4
Respect for Others 4
Cost/Benefit 4

Science in Business Strategy 5

Current Performance Levels in the Business Community 7

Common Blunders in Business Strategy 8
Making Moves Inappropriate to the Position 8
Building Productivity in the Wrong Way 9
Underdeveloping Strong Positions ... 9
Underdeveloping Weak Positions 9
Going Down With the Flagship 9
"Zigzagging" 9
Joining the Other Lemmings 9
Inappropriate Turnaround Efforts 10

Common Blunders in Logic and Method 10
Gullibility 10
Failure to Test Consequences 10
Poor Forecasting Logic 10
Wrong Peer Group as Basis for Evaluations 11
Tunnel Vision 11
Erroneous Discounting 11
Cost/Benefit Errors 11

Conventional Wisdom vs. Science 11
Investing in Technology 12
  The "Unprofitability" of Modern Technology 12
  What to Do About Investment-Intensive Requirements 17
Pursuing Market Share 19
  Conventional Wisdom vs. Observed Behavior 20
  Why Do Businesses Act Counter to Their Beliefs? 22

Conclusion 26

FIGURES

4-1 Investment Intensity and Return on Investment 12
4-2 Investment Intensity and the Ratio of Residual Income to Sales 13
4-3 Effect on Return on Investment of Investment Intensity and Market Share 14
4-4 Increases in Selling Prices Related to Investment Intensity 15
4-5 Value Added Related to Investment Intensity 16
4-6 Breadth of Product Line Related to Investment Intensity 19
4-7 Adequate Productivity Improvement for Increased Investment 19
4-8 Market Share Gains 20

4-9 Effect of Real Market Growth on Share Gain 21
4-10 Effect of Initial Share Level and Market Growth Rate on Change in Market Share 22
4-11 Cash Flows of Steady Market-Share Businesses vs. Gained Market Share Businesses 23
4-12 Returns on Investment of Steady Market-Share Businesses vs. Gained Market-Share Businesses 24
4-13 Cash Costs of Gaining Market Share 24
4-14 Effect on Return on Investment of Gaining Market Share 25

WHY EXAMINE THE ISSUE?

In most areas of human productive activity there is now, near the end of the twentieth century, absolutely no need to inquire what role science can and should play in that activity. The basic answer is obvious. Almost everyone understands that without science a modern doctor would be no better than a primitive medicine man, a modern engineer no more capable than a medieval artisan, and a modern farmer no more productive than a stone age hunter and gatherer. There are many issues on details and specifics (and always will be), and moral questions on use vs. abuse (and also always will be), but no disagreement on the basic notion that the laws of nature are capable of being discovered, important to know, and employable for the benefit of man. Moreover, the technology for making discoveries and applying them to human purposes is well developed, extensively tested, continually being further improved, and very extensively employed.

Curiously, this is not so in the area of business management, particularly strategic business management. The most popular concept here is still that business management is a craft. Strategic management is frequently seen as something that is learned by direct experience or from another craftsman, that consists of little more than the application of common sense plus native intelligence to a particular kind of problem, and that can be codified in relatively simple and unchanging procedures. In an increasing number of industries, one can see the remarkable spectacle of the product becoming better and better and the process of production becoming more and more efficient, while the business and its management remain qualitatively stagnant. The situation is very similar to military strategy, where the quality of the weapons keeps improving, but the quality of the generals remains the same.

How can science improve the situation? This is not the place for a treatise on scientific method; in any case, that would amount to pontification about the well known. Instead, the question is approached from the negative side: What business blunders are occurring because science is not used in business strategy, and how might these blunders be minimized? A summary of the key concepts helps to clarify the major points.

WHAT IS "SCIENCE"?

"Science," as used here, does not mean test tubes or telescopes, but refers to a method of thought. This method has proven exceedingly powerful in all areas of human endeavor to which it has been applied; it produces insights and capabilities that are cumulative, expanding, and improving without end and without upper limit. Progress is not, of course, always free of misstep or error (or even fraud), but part of the essential nature of science is that these errors are discovered and corrected.

Some of the key ingredients of this method of thought are discussed in the following sections.

Observation and Evidence

One learns about the world by observing the world. Likewise, one learns about business strategy by observing business strategy. Observation must be careful rather than sloppy: The observer measures, records, compares, and checks. All opinions and theories are provisional until they are verified by adequate evidence, and all evidence consists of careful observation of the real world. If the real world is kind enough to show the observer what he or she wants to see, then it is observed; otherwise the observer tries to coax the world to reveal its secrets by arranging suitable experiments.

Analytical Rigor

Evidence is interpreted through the use of logic. One does not jump to conclusions, but works toward them. When an inspiration or a flash of insight does occur, it is necessary to go back and check to see if the insight is logically consistent with the evidence and if the evidence is adequate. If the idea is inconsistent with the body of available evidence, it is rejected as erroneous; if it is consistent with the evidence but the evidence is inadequate, judgment is reserved until the evidence is adequate. The standards of adequacy are tight rather than loose.

Significance

When a pattern or relationship is observed, it must be tested for significance before it is accepted as a "fact" or "law." Thus the risk that the observed relationship is accidental and ephemeral rather than systematic and dependable is minimized. If an observed pattern is not significant, little credence is placed in it. If it is not known whether the pattern is significant (perhaps because its significance has not been tested), the pattern is not considered significant and little credence is placed in it.

Probability

Patterns and relationships are not expected to be black or white; they can be shades of gray. For example, cigarette smoking does not always cause lung cancer, and not smoking does not always prevent it, but the probabilities are different (and the differences are significant). The important thing is to measure the probabilities. Unknown probabilities are poor guides to action.

Hypothesis and Test

Knowledge accumulates by a cyclical and repetitive process of hypothesis and test. The hypothesis phase is the driving force; the test phase keeps it honest. Imagination, creativity, energy, hopes, and dreams come into play during the hypothesis phase; conscience, caution, and soundness come into play during the test phase. It is important that both phases be kept active all the time. (In this context, "theory" is an advanced form of hypothesis.)

Representative Cross Section

Conclusions with respect to a particular case are almost always derived from similar other cases. The criteria for similar other cases are subject to hypothesis and test. A sufficient number of similar other cases must be examined in order to test for significance. Therefore, when searching for reliable action guides in any particular situation, the trick is to find an adequate number of other cases that are similar to the present case. Failing that, there are no reliable action guides.

Respect for Others

Evidence is assembled and conclusions are reached in a way that can be communicated to others and verified by them. This minimizes the risk of wishful thinking, error, or fraud. It also minimizes the wastage that comes from reinventions of the wheel. The habit of communicating both evidence and logic is a manifestation of respect for others. The conclusions or recommendations of those who are unwilling or unable to support their views are rejected.

Cost/Benefit

The principles of scientific observation may be deviated from for clear cost/benefit reasons. Sometimes it is simply too much trouble to do things right;

for example, where the cost of error is small. When in doubt, however, all the principles should be applied in a conscientious and disciplined manner.

SCIENCE IN BUSINESS STRATEGY

The foregoing does not, of course, constitute a rigorous definition of science or a complete description of its methods. Emphasis has been placed on those precepts that, even though they are of great importance, are very frequently neglected in the formulation of business strategy and policy.

Strategic issues are, by definition, the critically important issues in a business—those that spell the major differences between success and failure. Therefore, they clearly deserve the best thinking and the most powerful methods and tools that can be brought to bear upon them.

Why, then, is the nonscientific approach to strategy-setting widely prevalent? Probably the most important fact is that the scientific base for business strategy is in a regrettably underdeveloped state and simply does not suffice as a guide to many of the decisions that must be made. It is not that the scientific approach is in principle unable to cope with business problems; it is that use of the approach has not been widely attempted. There is a vicious cycle at work here: The "supply" of the fruits of science is low, so the "demand" is low, so the "supply" is low, and so on.

The problem is how to break this vicious cycle and to turn it into the kind of benign cycle that works so effectively in medicine, engineering, and agriculture. Fortunately, some very good beginnings have already been made.

Chief among these, in the specific area of business strategy, is the Profit Impact of Market Strategy (PIMS) Program. This program, organized at the Harvard Business School in 1972 and operated by the nonprofit Strategic Planning Institute since 1975, is a larger-scale continuation of previous work done at the General Electric Company. It is, in effect, a cooperative venture, involving about 250 major companies in North America and Europe, designed to provide a set of observations, plus associated analytical procedures, for a more scientific approach to the planning efforts of the participating companies.

Each member company contributes information about its experiences in several different business areas to a combined data base. The PIMS staff analyzes this experience to discover the general laws that determine what business strategy in what kind of competitive environment produces what results. The findings are made available to member companies in a form useful to their business planning.

The program provides business managers and planners with tools and data to answer questions such as:

- What profit rate is normal for a given business, considering its particular market, competitive position, technology, cost structure, and so on?
- If the business continues on its current track, what will its future operating results be?
- What strategic changes in the business have promise of improving these results?
- Given a specified contemplated future strategy for the business, how will short-term and long-term profitability or cash flow change?

In each case, the answers are derived from an analysis of the experiences of other businesses operating under similar conditions.

More specifically, four purposes of the program are:

1. To assemble a data base reflecting the business strategy experiences of a group of participating companies;
2. To conduct a research program on that data base in order to discover the laws of the marketplace that govern (a) profit levels, (b) other outcomes of strategic actions, and (c) outcomes of changes in the business environment;
3. To conduct an applications program to make the findings of the research available to participating companies in a form and manner they can use effectively; and
4. To carry-out ancillary activities (publication, education, service to participants, study of planning methods, legislative recommendations) that will enhance the value of the entire program to its members and to the economy at large.

The unit of observation in PIMS is a business. Each business is a division, product line, or other profit center within its parent company which sells a distinct set of products and/or services to an identifiable group of customers in competition with a well-defined set of competitors and for which revenues, operating costs, investments, and strategic plans can be separated in a meaningful way.

Currently the data base consists of information on the strategic experiences of more than 2,500 businesses, covering a three- to ten-year period. The information on each business consists of about 100 items descriptive of the characteristics of the market environment, the state of competition, the strategy pursued by the business, and the operating results obtained. Each data item has been pretested for significance and relevance to profitability.

The PIMS staff has devised a set of standardized forms to be filled out by the participant company for the contribution of its experience records to the data bank. The forms are designed to break the required data items into simple elements that can be assembled readily from financial or marketing records, or that can be estimated by someone familiar with the specific business.

One of the most encouraging aspects of the PIMS Program is the growing volume of academic research being performed on the PIMS data base. Numerous studies by academicians (faculty research and doctoral dissertations) have been completed. These studies, plus the increasing amount of other scientific work now being done at such institutions as Harvard, Columbia, Purdue, Michigan, Stanford, and New York Universities, are providing a demonstration to both the business community and business schools that substantial progress is indeed possible in this area.

The fruits of the effort will accumulate to the point where most decisions in the areas of business strategy and business policy may become as professional as corresponding decisions in, for instance, medicine. In the meantime, what is the practitioner to do? Life must go on, whether science is fully ready to assist or not.

The first thing to do is, of course, to take full advantage of the findings and the tools already available. They do not solve all problems, or even most, but they perform quite well as far as they go. Second, the effort to keep improving the data, the research, and the tools should receive support. Third, a sense of discrimination between the "good stuff" and the makeshifts is needed; in other words, the practitioner must become a more demanding consumer of business technology.

The field of health care, in addition to offering much instruction on how to become more scientific, also shows what can be done where the benefits of science have not yet reached. Where a disease can be cured, it is. Where enough is not yet known, (a) the patient is made as comfortable as possible; (b) his mind is occupied with good thoughts or distracting activities; (c) a miracle is hoped for; (d) if necessary, the process of dying is eased; and (e) work continues to do better the next time. A well-designed process of strategic planning does the same for the business executive.

CURRENT PERFORMANCE LEVELS IN THE BUSINESS COMMUNITY

The PIMS data base, in addition to providing observations for a scientific approach to business strategy, also affords a view of the damage done by the failure to use this approach.

The PIMS Program conducted a study to measure the shortfall from potential performance in a sample of 120 businesses. The approach took the following three steps:

1. For each business in the sample, measurements were made of the results of a "no-change" strategy (i.e., the operating results that the business will probably achieve if it simply maintains its current stra-

tegic posture and floats with the market). Operating results were measured in terms of the discounted present value of net cash flows plus capital gains (i.e., increase or decrease in the market value of the business) taken over a five-year period.

2. The potential operating results were then measured, using the same measure of performance. The potential was defined in terms of the results obtained by the good performers among "strategic look-alikes" (i.e., other businesses operating under strategically similar circumstances: having a similar position, in a similar market, facing similar competition, and employing a similar technology). The good performers were taken at approximately the eightieth percentile among the strategic look-alikes.
3. By direct comparison, a calculation was made for each business of its shortfall (if any) from potential.

The conclusions of this analysis certainly were not exact or indisputable, but they did produce a rough indication of the magnitude of the problem. The average business in the sample attained about half of its potential performance. This estimate of the current wastage of economic resources is in some respects too high (there is such a thing as simple bad luck, since the world operates on a probability basis, not a mechanical basis) and in other respects too low (the sample consisted of more outstanding companies than mediocre companies, and only of their surviving businesses). However, there is little doubt that the possibility of improvement is huge and that a scientific approach can make most of the difference.

COMMON BLUNDERS IN BUSINESS STRATEGY

Some of the specific mistakes that lead to the waste of resources can be identified. This discussion does not concern unlucky moves, but those that can clearly be expected, on the available evidence, to fail.

Making Moves Inappropriate to the Position

Many businesses make moves and take actions that have a high failure rate for the starting position of the business. Examine the following examples of inappropriate strategies:

- XYZ, Inc., a business with a low relative market share, simultaneously carries out an expensive and innovation-oriented research and development program and an expensive and leadership-oriented marketing program.
- Continental Flange, a business with a product of mediocre quality, attempts to compensate for lack of quality by an energetic marketing effort.

- R.P.H., Inc., a business with a low relative market share, becomes vertically integrated.

Building Productivity in the Wrong Way

Many businesses, in an effort to improve the productivity of their work force, mechanize or automate excessively, increasing their fixed costs considerably. As a result they stand to lose the fruits of their improved productivity by becoming embroiled in frequent and severe price wars or marketing wars.

Underdeveloping Strong Positions

Many well-positioned businesses (good product quality, well-segmented market, high productivity) tend to rest on their laurels and fail to take full advantage of the opportunities (e.g., for backward integration or for market-centered diversification) that their position provides.

Underdeveloping Weak Positions

Many businesses whose current positions have a few elements of weakness (e.g., slow market growth or low relative share) but other elements of strength give up prematurely and underinvest in the business by a considerable amount.

Going Down With the Flagship

Many companies depend heavily on a major "flagship" business. When confronted with maturity in that business or with a major new competitive challenge (e.g., from Japan), they fail either to restructure the business to cope with the new realities or to redeploy their strengths to other businesses or markets.

"Zigzagging"

Lacking the fortitude and clear-sightedness to see a strategy through to a successful conclusion, many businesses change position or strategy frequently, dissipating resources through wasteful "zigzagging."

Joining the Other Lemmings

Joining the rush into apparently promising new markets or technologies without adequate competitive strengths to succeed in these new areas, a business may well be unable to survive the inevitable shakeout.

Inappropriate Turnaround Efforts

When a business gets into profit trouble or cash trouble, it may predicate its strategy on moves that produce only a short-run appearance of improvement at the cost of an erosion of the fundamentals of the business. A common example is saving money by reducing product quality.

COMMON BLUNDERS IN LOGIC AND METHOD

The blunders in business strategy are attributable, in a general way, to shortcomings in logic and method. But under that broad heading, there are some specific errors that appear with particular frequency and cause an unusual amount of mischief.

Gullibility

Gullibility involves an uncritical acceptance of simplistic and/or unsupported principles of strategy and rules of business conduct. At least half of the most frequently cited principles of business strategy are not grounded in any visible evidence, or are inappropriate in major groups of businesses. It is easy to tell the sound principles from the unsound ones by examining the supporting evidence. If there is no evidence, or if the evidence is inadequate, the alleged principle is probably untrustworthy.

Failure to Test Consequences

Another common mistake is the adoption of a strategy for a particular business without a test of the consequences of that strategy. Many companies choose a strategy simply on the basis of where the business fits on a matrix, what their competitors are doing, what an expert recommends, or the prevailing consensus of the management team, without taking the trouble to test that strategy by a credible forecast of its consequences. It is not always easy to make this forecast, but it is always important to make the effort.

Poor Forecasting Logic

Inappropriate forecasting models are often used in strategy planning. The most frequent error is to employ time-series-based longitudinal models for strategic forecasting. This is almost always an error, because any change in the basic structure of the business or its environment will invalidate a longitudinal model, and strategic changes normally do involve the basic structure. The answer is to use models with cross-sectional bases, whose range of applicability spans both the old structure and the new structure.

Wrong Peer Group as Basis for Evaluations

A business may use an inappropriate group of observations to evaluate its strategic options. Many business people compare their own situation to that of other companies in their industry. This is usually misleading, because other companies do not always share the same strategic problems. For example, the passenger car business of General Motors has little to learn from the experiences of Ford, Chrysler, or Toyota with passenger cars, but a lot to learn from the IBM mainframe business or the General Electric lamp business, which also hold top rank in a maturing industry under severe competitive attack.

Tunnel Vision

This error consists of exploring too narrow a range of strategic options, thereby missing opportunities for substantial (rather than merely incremental) improvements in the business.

Erroneous Discounting

In evaluating future incomes or outgoes, businesses may use interest rates that are too high (thereby condemning the business to a harvest strategy) or too low (thereby ignoring an important element of opportunity cost). The villain is often inflation, which confuses many people as to which interest rate is really applicable. Most discount rates of less than 2 percent or more than 5 percent in real terms (i.e., relative to the rate of inflation) are suspect and probably wrong, unless the risk of the investment is very high. The use of a very short planning horizon (less than five years) is tantamount to a too-high discount rate.

Cost/Benefit Errors

Overinvesting or underinvesting in the planning effort, relative to the benefits achievable, is another common blunder in planning. There is a frequent inverse correlation between the cost and the benefit of a planning process. This happens because an inadequacy of logic or method is often compensated for by elaborate bureaucratic busywork or expensive consultant studies. Good planning is not very expensive, compared to bad planning, but it is not cheap either, because it must be done with considerable care and attention to possible missteps and errors.

CONVENTIONAL WISDOM VS. SCIENCE

Science often challenges conventional wisdom, with the outcome being improved potential for effective health care, space exploration, data processing, corporate performance, and so forth. The following examples illustrate the role of science in challenging conventional wisdom in business strategy.

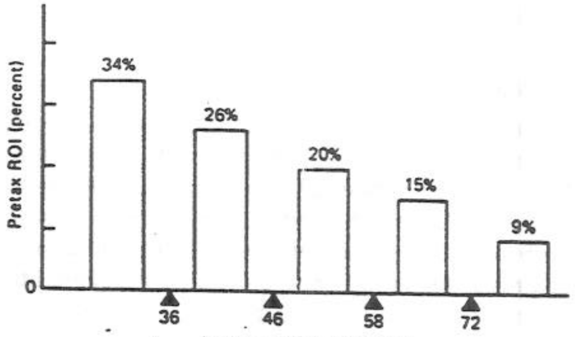
Investing in Technology

The "Unprofitability" of Modern Technology. Of all the findings on business strategy yielded by the study of the more than 2,500 businesses in the PIMS data base, the following remains one of the most controversial: Businesses that are highly investment intensive, that is, those that use high levels of investment per dollar of sales revenue (airlines, bulk chemical-processing plants, or distributors of consumer goods requiring large inventories), are much less profitable than businesses with lower levels of investment per dollar of sales. Figure 4-1 illustrates this phenomenon in terms of return on investment (ROI).

This finding is controversial not because the phenomenon is rare, uncertain, or weak—it is common, quite clear, and extremely powerful—but because it is so unexpected. The conventional wisdom is that there is a strong positive relationship between investment intensity and modernity or progressiveness. Everyone knows that modern technology requires elaborate machinery, and thus heavy investments; that high labor productivity depends on extensive automation, and thus on heavy investments; and that

FIGURE 4-1

Investment Intensity and Return on Investment



Note: "Investment" includes both working capital and fixed capital (at net book value). The figures are four-year averages. The cut points used have no significance in themselves, but are chosen to divide the businesses in the data base into equal groups. With a different definition of investment (e.g., predepreciation book value or current replacement value) or a different performance index (e.g., profit as a percentage of value added) or a different time period (e.g., three years or five years), the specific numbers are moderately different, but the strategic message is the same.

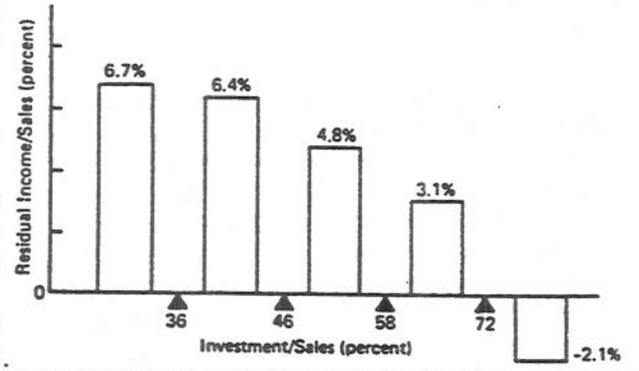
consumer goods must be readily available to customers to sell successfully, and thus require large inventory investments. Since modern technology, high labor productivity, and readily available consumer products are judged to be good things, they are expected to improve profitability rather than to hurt it. What in fact happens is that the commonly expected public benefits of investment-intensive technology (higher wages, lower prices, and improved product quality) do indeed occur most of the time; but, alas, the expected private benefits do not occur. Instead, the profits of companies that use an investment-intensive technology are usually rather poor.

Why does investment intensity hurt profits? Before answering this question, it is important to confirm that we are in fact dealing with an issue that is real and substantive, not just an optical illusion. For example, it might be argued that the negative pattern in Figure 4-1 is due entirely to arithmetic: that investment-intensive businesses have large denominators in their ROI ratios, and that their returns are low for that reason alone. That possibility can be excluded by using another measure of profitability. For example, Figure 4-2 relates investment intensity to the ratio of residual income to sales. The similarity of the pattern to that of Figure 4-1 confirms that investment-intensive businesses are actually less profitable.

Another possible argument is that the negative effect of investment intensity on profit is actually due to something quite different. Specifically, it is well known that businesses with high shares of their served markets are considerably more profitable than those with low shares. One of the reasons for that relationship is that high-share businesses tend to be more efficient

FIGURE 4-2

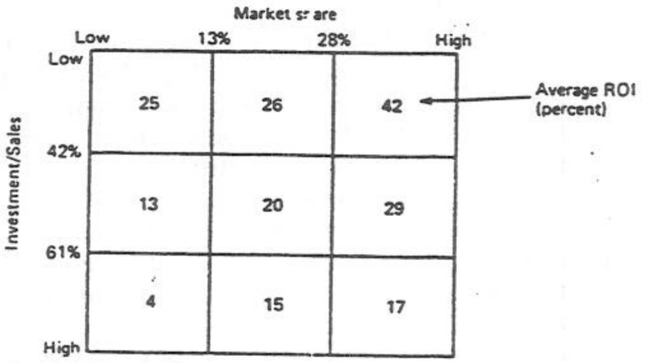
Investment Intensity and the Ratio of Residual Income to Sales



Note: "Residual income" equals pretax operating income less a financial charge of 10 percent on the investment used in the business. Other rates of financial charge produce different slopes but the same basic pattern. Even so unrealistic an assumption as a zero cost of money produces a slightly negative slope.

FIGURE 4-3

Effect on Return on Investment of Investment Intensity and Market Share



Note: The cut points between adjacent rows and columns were chosen to give approximately equal numbers of businesses in each row and each column. The ROI number in each box is the four-year average for the businesses falling into that box.

users of investment and are therefore less investment intensive, while low-share businesses are often inefficient users of investment and are therefore more investment intensive. The negative effect on profit of high investment intensity might therefore be no more than the reverse side of the coin of the favorable effect on profit of high market share. To exclude that possibility, it is necessary to establish that there is an effect on ROI from investment intensity in addition to that from market share. Figure 4-3 shows that there is indeed such an effect.

All the businesses in a given column have similar market shares, but investment intensity increases with each row. The fact that ROI decreases sharply moving down each column confirms that investment intensity has a negative effect on profit, regardless of market share.

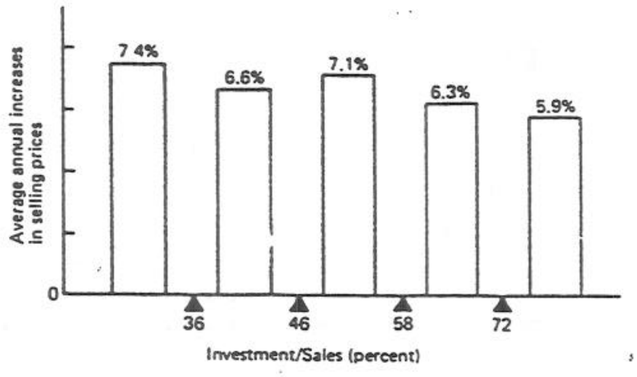
Within each of the classes of investment intensity, one can examine not only ROI itself but also the degree to which the ROI of the group of businesses is lower or higher than would be expected from their levels of market share alone. Investment-intensive businesses are found to be even less profitable than would be expected from their market share, further confirming the phenomenon as a real and important one.

A third argument is that low ratios of sales to investment (i.e., high levels of investment intensity) are actually due to low levels of capacity utilization. A similar test of this hypothesis yields a similar result.

The major reason for this negative effect seems to be that the game of competition is played in a very different way in investment-intensive industries than in others. When each of the firms competing in a partic-

FIGURE 4-4

Increases in Selling Prices Related to Investment Intensity



ular industry has committed heavy investments on which a reasonable return needs to be earned, each becomes rather eager to keep its capacity loaded. In an investment-intensive facility, volume is commonly believed to be the key to profitability. (Perhaps surprisingly, this belief is as common in industries where the investment consists largely of working capital as it is in those where the investment is largely fixed capital.) So the competitive process in investment-intensive industries readily degenerates into a volume-grubbing contest, punctuated with frequent price wars, marketing wars, and other over-intensive competitive measures that take most of the joy out of being modern, automated, or otherwise investment intensive. In particularly good years, when every company's capacity is almost fully loaded, this effect may not appear at all, but when good years are averaged with bad, as in the figures presented here, the negative effect on profit is quite clear.

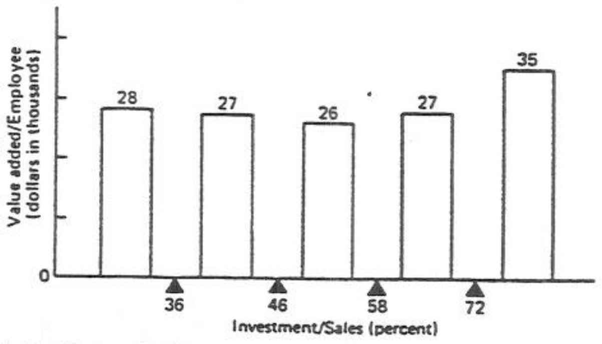
Figure 4-4 illustrates this point by relating increases in selling prices to investment intensity. The contrasts are most obvious at the extremes.

The profit-depressing price squeeze also shows up clearly in examining the value added per employee in industries that differ in their investment intensity. "Value added" is the degree to which a business upgrades the market value of the raw materials or components it buys, i.e., the difference in market value between what the business buys and what it sells. Figure 4-5 shows that the value added per employee does not increase with added investment intensity over a surprisingly wide range.

Only on the extreme right of Figure 4-5 is there any significant favorable impact; in four fifths of all businesses, it appears that employees working with the support of large investments produce no more value than those

FIGURE 4-5

Value Added Related to Investment Intensity



Note: The estimate of "value added" has been adjusted for abnormally high or low earnings in particular industries.

working with only small investments. How can this be? It is possible that the first group of employees is simply less productive as individuals, that they are less capable people. More likely, however, is that the more intensive price competition occurring in investment-intensive industries reduces the market value of the product of companies in those industries, and hence the value added by their activities. Even on the far right of Figure 4-5, the increase in value added per employee is usually insufficient to justify the higher investment and cash required, as verified by examining the profit and cash results of most of these businesses.

A closely related reason for the unprofitability of investment-intensive businesses lies in their apparent inability to manage their personnel levels in a suitable fashion: The number of employees per dollar of sales is just about as high as in less investment-intensive industries. The increased numbers of people in administrative, commercial, and other functions on average counterbalance the savings of people in manufacturing functions.

Before leaving this subject, it is worth noting that Figures 4-1 through 4-5 have undoubtedly slanted reality in favor of investment-intensive businesses, and therefore have understated the magnitude of the problem. This bias occurs because the statistics are drawn from accounting records that are kept in the conventional way. It is becoming widely recognized that conventional accounting procedures overstate ROI, particularly of investment-intensive businesses, both by overstating the numerator and by understating the denominator. The reason is the same in both cases: Fixed assets are valued at their historical cost, rather than at their current inflated worth, understating both depreciation costs (and thus increasing apparent earnings) and investment levels.

What to Do About Investment-Intensive Requirements. If the technology of a business clearly requires a high level of investment intensity, either of plant and equipment or of working capital (or of both), what actions should the business take?

First, the business should not automatically assume that more is better insofar as investment intensity is concerned. A highly automated plant is not necessarily a better plant than a less automated one. Of course, the recent series of cash crunches in the economy has already cured many businesspeople of this knee-jerk reaction, but old habits of thought die hard, particularly in industries, such as bulk chemicals, where increased capital intensification has long been a way of life. Most businesspeople still unquestioningly assume that good technology is synonymous with automated production or long and full pipelines in distribution. That assumption is valid just often enough in particular cases to save it from being ludicrous on its face. However, the proverbial "hard look" is clearly justified with regard to all investments that are larger than merely proportional to an increase in capacity.

Second, in evaluating a proposed investment that is clearly larger than merely proportional to an increase in capacity, the business should consider the strategic effect as carefully as it does the cost effect of the project. A negative strategic effect may more than offset a positive cost effect.

Suppose, for example, that a manufacturer is considering a capital-intensifying project (e.g., an increase in the degree of automation of plant) that on the basis of conventional cost calculations will have an annual operating cost, everything included, of $1 million, an annual saving of $3 million in reduced labor costs or reduced spoilage, and therefore a net benefit of $2 million per year. Suppose further that all of these estimates are absolutely accurate, and that the net saving of $2 million per year does in fact materialize as promised. Even then, the project may not be, and very frequently is not, profitable. While the business saves $2 million, it does not get to keep that $2 million. Instead, it gives the $2 million away, mostly to its customers, secondarily to its employees, and thirdly in the form of higher marketing costs.

A business is forced into such a move by the changed competitive climate. First, its competitors' and its own increased desperation for volume (to keep those expensive plants with their high fixed costs loaded) leads to price and other concessions to customers, particularly after the new technology has become widespread; second, increased fear of plant shutdowns leads to higher wage settlements and a greater reluctance to discharge unneeded people; and third, the business becomes caught in more and more intense marketing wars. So, frequently the net effect of the new technology is that its benefits accrue to the customer and to the labor union, while the business is left with the costs and the investments.

Now, the message of this scenario is not, of course, that any capital-intensifying investment should automatically be rejected, but rather that its strategic effect should be estimated as carefully, before the event, as its cost effect. The net result in many cases will clearly be against the contemplated project, no matter how glamorous it may be. However, while the project may have a negative net prospect as far as percentage ROI is concerned, it may quite often have a favorable prospect as far as dollar results are concerned, because the lower percentage is applied to a larger investment base. In such cases, a rather difficult management decision is required—difficult because it involves a tradeoff between an increase in sales and dollar profits on the one hand and a decrease in the rate of profitability on the other.

Third, the business should adopt a market strategy that minimizes the profit-damaging effect of capital-intensive technology. The PIMS data base suggests several ways of accomplishing this, many of them leading to increased dollar earnings, if not increased ROI levels or higher percentages on sales. Three of these moves are examined here.

The first is suggested by Figure 4-3. The bottom row, where the highly investment-intensive businesses are located, reveals a sharp increase in profitability moving from left to right, from 4 percent average ROI to 17 percent. Even at approximately equal levels of investment intensity, the high-share businesses do much better than the low-share businesses. This observation may seem rather obvious and unhelpful, until it is recalled that market share is measured relative to the served market, that is, relative to that segment of the total potential market in which the business is making a serious competitive attempt. So one way in which a business can obtain high market share is to concentrate its efforts on a segment of the total potential market—for example, a smaller geographic area or a more specialized class of customers. Since market segmentation efforts often can move a business toward the right on Figure 4-3, market segmentation or redefinition is clearly one strategic answer to the profit-depressing effects of capital intensity. The recent history of some segments of the specialty steel and pharmaceutical industries illustrates this principle.

A second approach is implied by Figure 4-6. Here businesses are divided according to their investment intensity and according to whether the breadth of their product line is narrower, about the same, or wider than what competitors are offering (to the same served market).

Focusing on the investment-intensive businesses in the bottom row shows that a product line that is broader than competitors' is clearly preferable to one that is not. This observation can be combined with the previous one: An investment-intensive business can be quite profitable if it focuses on a relatively narrow and conquerable market segment, and covers that segment really well, with a broad and diverse product line tuned to the preferences of that segment.

FIGURE 4-6

Breadth of Product Line Related to Investment Intensity

Investment/Sales	Relative breadth of product line Narrower		Wider
Low – 42%	27	29	33
42% – 61%	17	18	25
61% – High	7	12	15

ROI (percent)

A third approach to profit protection in investment-intensive businesses is to be sure to obtain adequate productivity improvement for the increased investment. Productivity rises more rarely than one would expect, but where it does rise, it certainly helps. Figure 4-7 gives the key facts.

Pursuing Market Share

According to conventional wisdom, a business with a large share of its served market, especially if that served market is growing rapidly, not only

FIGURE 4-7

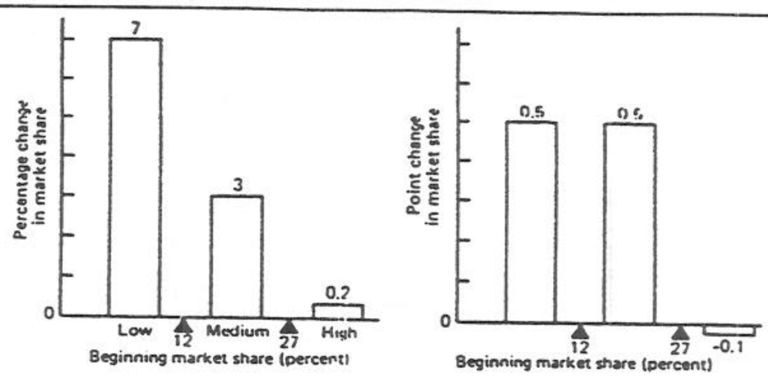
Adequate Productivity Improvement for Increased Investment

Investment/Sales	Value added/Employee Low – 19K	19K – 28K	28K – High
Low – 42%	23	29	37
42% – 61%	20	19	25
61% – High	7	10	15

ROI (percent)

FIGURE 4-8

Market Share Gains



Definitions:

Market share any year	Market share previous year	Percentage change in market share	Point change in market share
12	10	+20%	2
30	32	-7%	-2

Notes:

(1) The market-share cut points were chosen to divide the data base into equal groups, and have no significance in themselves. Each group in the figure represents over 1,000 business experiences.

(2) The average market-share gain for the economy at large would seem to be zero, since whatever share one company gains another loses. The tables here do not average to zero because (a) more "good" than "poor" businesses are represented in the PIMS data base and (b) some businesses are implementing a segmentation strategy, in which almost everyone in the industry has a rising market share (though of a diminished served market).

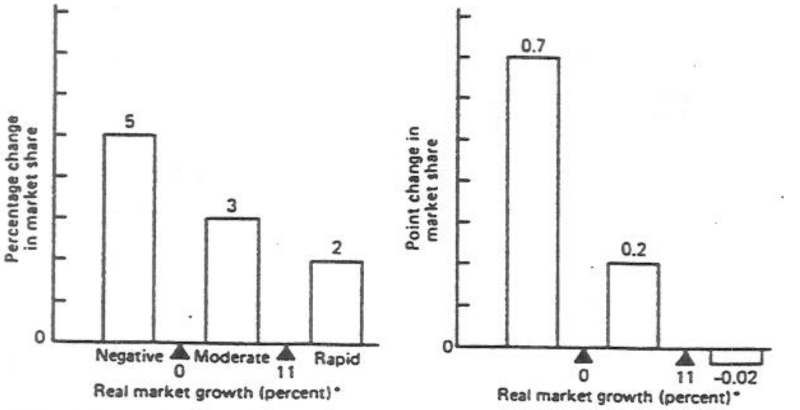
finds it easy to gain additional market share, but also should pursue a share-gaining strategy. Many businesspeople believe that large-share firms possess advantages in resource availability, strategic position, and visibility—all of which facilitate a strategy of trying for an even greater market share. Rapid growth, meanwhile, helps to reduce the cost of the move, because the risk of triggering a major competitive donnybrook is lower in a growing market. It is also a common belief that weak-share businesses located in low-growth markets should, and usually do, harvest some of their share. The PIMS models in most cases confirm these judgments as to how businesses should act. But, the PIMS data base also indicates that this is not how businesses in fact do usually act.

Conventional Wisdom vs. Observed Behavior. Businesses that already have high shares of their markets tend, on average, to gain less additional market share per year than do low- and medium-share businesses. Figure 4-8 shows this key fact in terms of both annual percentage changes and annual point changes in market share.

While one cannot conclude from Figure 4-8 that large-share businesses

FIGURE 4-9

Effect of Real Market Growth on Share Gain



**Real market growth*—Annual percentage growth of the served market in constant dollars.

are or are not able to gain share more easily than small-share businesses, it is clear that large-share businesses have not gained the most share and have not protected their leading market positions. Weak-share businesses, in contrast, averaged a gain of more than a half point of share, an increase of over 7 percent from their initial position.[1]

Also contrary to popular belief is the fact that the highest market-share gains occur when the short-run rate of market growth is negative. The least amounts of share gain occur in rapid-growth markets. This relationship can be seen in Figure 4-9 (where, as before, each group represents well over 1,000 business experiences[2]).

To what extent are share gains in declining markets due to exit by some competitors? While some minor improvement in share performance does seem to be due to exits from the industry, the growth of market share of the survivors is greater than can be accounted for by this phenomenon alone.

Large-share businesses, on the average, lose share in rapid-growth environments, while weak-share firms gain share in such environments. Equally surprising, large-share businesses fare worse than weak-share businesses in negative market-growth situations. Figure 4-10 indicates the combined

---

[1] The result is the same if initial relative market share is substituted for the initial simple market share.

[2] The PIMS data base has five or more years of data for over 1,000 businesses. By taking every year-to-year move as a separate observation, over 4,000 year-to-year experiences are available for examination.

FIGURE 4-10

Effect of Initial Share Level and Market Growth Rate on Change in Market Share

Real market growth

Beginning market share (percent)	Negative 0%	0% 11%	11% Rapid
Low 12	10	7	5
12 27	6	3	2
27 High	2	0	-1

Percentage change in market share

Real market growth

Beginning market share (percent)	Negative 0%	0% 11%	11% Rapid
Low 12	0.8	0.5	0.4
12 27	0.9	0.4	0.2
27 High	0.5	-0.1	-0.5

Point change in market share

impact of initial share level and market growth rate on change in market share. It was the large-share businesses in rapid-growth markets that the conventional wisdom suggested would achieve the greatest share gains.

Why Do Businesses Act Counter to Their Beliefs? The focus here is on the large-share businesses, since they act in the most surprising manner. The observed results, counter to expectations, occur either because there really are natural market forces at work that lead such businesses to lose or give up share in rapid-growth markets, or because these businesses choose to follow that path. Among the reasons why large-share businesses may choose not to be aggressive could be the high costs of pursuing such a strategy.

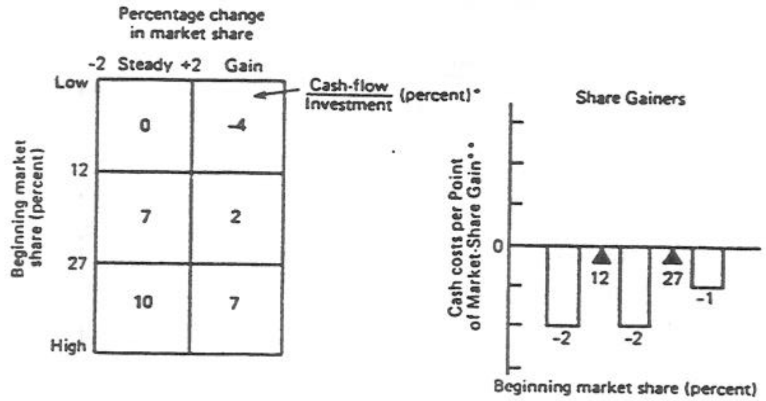
Is it sensible, from a short-term performance perspective, for large-share firms to hold market position steady rather than to build it further? This question is examined in Figure 4-11 by comparing the cash flow, and in Figure 4-12 by comparing the ROI, of businesses that held market share steady and businesses that gained market share.

Figure 4-11 indicates the short-term cash costs incurred in gaining additional share. For example, among large-share businesses (bottom row), the market-share gainers earned a cash flow of 7 percentage points, compared to 10 points for the businesses holding their positions steady. Thus, the overall relative cash cost to the share gainers was 3 percentage points. Cash costs are moderately, but not significantly, lower for large-share businesses than for small-share businesses, both overall (−3 vs. −4) and per point of share gain (−1 vs. −2).

But, for short-term ROI cost, the story is quite different. Large-share

FIGURE 4-11

Cash Flows of Steady Market-Share Businesses vs. Gained Market Share Businesses



$$^{*}\text{Cash flow/investment} = \frac{\text{After-tax net income} - \text{Increase in net investment}}{\text{Total investment}}$$

$$^{**}\text{Cash costs per point of share gain} = \frac{\text{Cash flow}_{(\text{Gainers})} - \text{Cash flow}_{(\text{Steady})}}{\text{Point change in market share}_{(\text{Gainers})}}$$

businesses must pay a short-run cost in terms of reduced ROI, while small- and medium-share businesses are able to increase short-term ROI and market share simultaneously. (Over a longer time horizon, however, large-share businesses will reap the ROI benefits of having a higher level of market share.)

Looking now at market-share change segregated by rate of market growth rather than by level of initial market position, conventional wisdom holds that it is least costly to gain market share in rapidly growing markets. The belief probably derives from the ability of all competitors to build their sales simultaneously in a growing market. Since no one's sales level needs to decrease, costly share-gain battles develop less frequently.

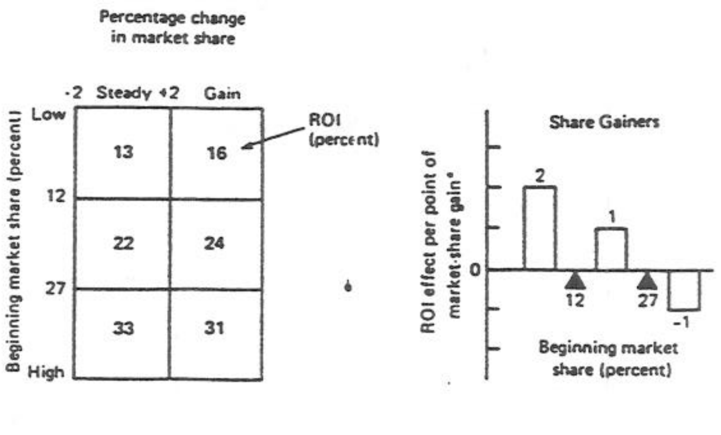
Cash and ROI performance, respectively, are illustrated in Figures 4-13 and 4-14.

Figure 4-13 indicates again that the process of building market share carries with it the cost of a lower short-term cash flow. Contrary to conventional wisdom, the cash costs of gaining share do not vary substantially between moderate- and rapid-growth markets. When cash costs are appropriately adjusted for average point change in market share in each environment, the cost of a point in share is only slightly lower in the more rapidly growing markets.

But, despite the modestly lower cash costs of building market share in

FIGURE 4-12

Returns on Investment of Steady Market-Share Businesses vs. Gained Market-Share Businesses



*ROI effect per point of share gain = $\frac{(ROI_{(Gainers)}\ \ ROI_{(Steady)})}{\text{Point change in market share}_{(Gainers)}}$

rapid-growth environments and the general attractiveness of having a strong position in a growth market, little share growth is actually occurring, as shown in Figures 4-9 and 4-10. Managers in such environments are usually not under severe pressure to actually gain share, since as long as they

FIGURE 4-13

Cash Costs of Gaining Market Share

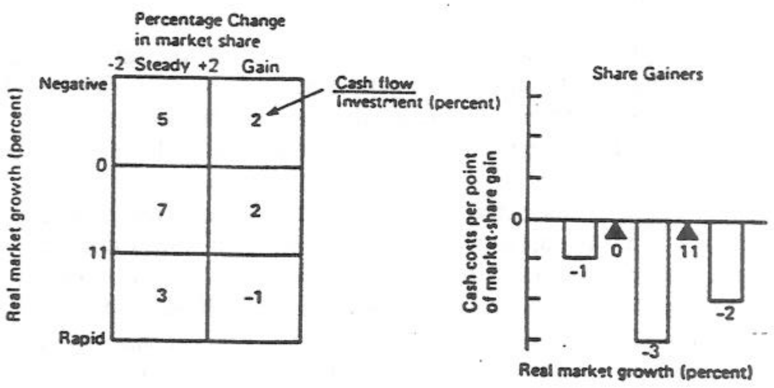
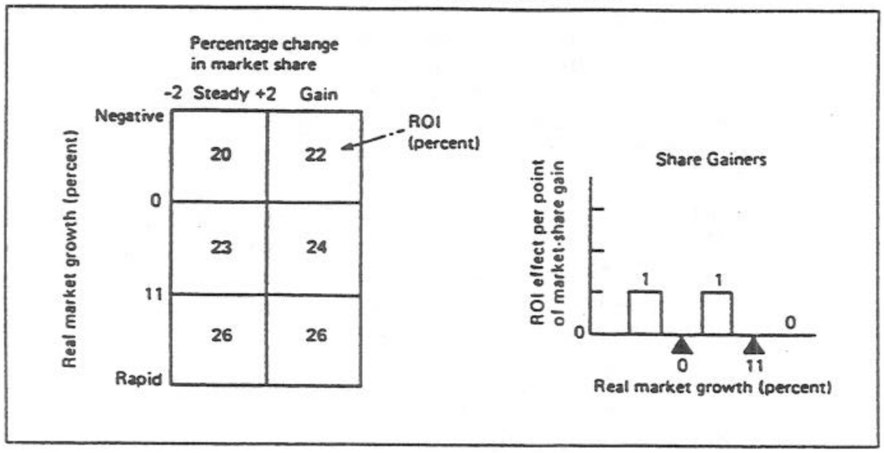


FIGURE 4-14
Effect on Return on Investment of Gaining Market Share



merely maintain share they will experience high profitability and significant sales growth.

Most surprising, perhaps, is the rapid share growth exhibited in moderate-growth markets—an environment where share growth is more costly and "harvesting" is frequently prescribed. Managers in this environment often turn to such tactics as creative product-line extensions to prolong the maturity phase of the product, and to build sales and market share simultaneously.

Businesses gaining market share do not pay a short-term ROI penalty regardless of the rate of market growth. The magnitude of the differences between those holding share steady and those building share, both with and without adjustment for the amount of share gain, does not allow a conclusion about whether building share costs more or less ROI in any particular market-growth environment.

While one can only speculate as to why so many people behave in a way counter to what they probably believe, the evidence suggests the following three explanations:

1. Managers have a short time horizon. Most managers are concerned with current results—profits this year and perhaps the next. They are encouraged to have this outlook by the fact that their own salaries, bonuses, and promotions may be tied to current performance. Thus, managers would be unwilling to suffer either cash costs or ROI costs now for higher share and higher ROI in the more distant future.
2. "Management by exception" is a management style that focuses upon things and events that are not behaving "normally" today.

This style, when applied to corporate strategy, may result in complacency about businesses that are very strong today, and encourage managers to focus instead upon bringing weak businesses up to average. From a portfolio perspective, such behavior may lead to investing in weak-share businesses that have lower ROI and cash-flow performance, rather than investing in strong-share, high-profit alternatives.

3. The large-share business may hesitate to use its share-gaining muscle out of fear of wreaking havoc in the market. Even a moderate expansionary move by a business with an already large market share could have quite substantial impacts on smaller competitors, and thus create quite a bit of market turmoil.

One conclusion: A strong corporate strategy is needed. Why? In the absence of such a strategy, it appears that natural business forces frequently push businesses in the wrong direction. Further, corporate strategy should be coupled with an appropriate reward system to induce managers to follow that strategy.

CONCLUSION

Good strategy planning consists of asking good questions, getting good answers, and effectively implementing those answers. The scientific approach is of major help in all three stages. It works because, on the positive, it increases the power of accomplishment and, on the negative, it minimizes the risk of costly blunders.

The scientific approach does have the drawback of a slow gestation period. But, while the accumulation of knowledge and method takes time, it is cumulative. The results get better and better, without any apparent upper limit. In any area of human activity in which the approach of science has ever been attempted, the step was always irreversible; it was never decided to go back to the old ways. There is every reason to expect that the field of business strategy and policy will travel the same course.

4-6

5

BCG

CORPORATE STRATEGY

The
Boston Consulting Group

The Boston Consulting Group, Inc.
One Boston Place
Boston, Massachusetts 02106
(617) 722-7800

The Boston Consulting Group, Inc.
200 South Wacker Drive
Chicago, Illinois 60606
(312) 993-3300

The Boston Consulting Group, Inc.
2180 Sand Hill Road
Menlo Park, California 94025
(415) 854-5515

The Boston Consulting Group, Inc.
333 South Grand Avenue
Los Angeles, California 90071
(213) 621-2772

The Boston Consulting Group, Ltd.
Bowater House East
68 Knightsbridge
London SW1X 7LP
(1) 581-1345

The Boston Consulting Group & Cie
Tour Fiat
92084 Paris la Defense Cedex 16
(1) 796-41-41

The Boston Consulting Group GmbH
Maximiliansplatz 5
8000 Munich 2
(89) 55 16 20

The Boston Consulting Group GmbH
Ürdingerstrasse 5
4000 Düsseldorf 30
(211) 45 10 63

The Boston Consulting Group K.K.
3-6, Ohtemachi 2-chome
Chiyoda-ku, Tokyo 100
(3) 279-0761

PERSPECTIVES

WHARTON REPROGRAPHICS

16

BUSINESS CHESS

The Boston Consulting Group

BUSINESS CHESS

Profound parallels exist between business and chess. Both are complex forms of competition. Both have been studied for centuries. and both depend on strategy.

Chess is a simplified. stylized representation of ancient conflict. Two forces of sixteen pieces. each in six types representing different rules of maneuver and engagement. are arrayed in perfect view of each other on an eight-by-eight playing field. The game lasts at most a few hours.

Yet even for this simplified representation of competition. the most powerful computer is unable to devise a winning strategy on the basis of analysis alone. Rather. sophisticated chess computers depend on rules of thumb and experience-based policies and procedures to develop strategies and direct tactics.

These rules to simplify a complex world are the parallels to the mental maps successful managers develop to determine a corporation's strategy and tactics. Like chess programs. these maps become implicit decision rules based on experience. Learning organizations build success on success as they revise and update their maps. Others continue to follow old maps and gradually become ineffective. Making maps explicit and continually reflective of reality separates grand masters from computer programs and successful businesses from bureaucracies.

The immense sophistication of business. compared with chess. makes us even more dependent on our experience and less aware of how much our maps guide behavior. Business is an infinitely more complex form of competition. The number of players is almost always more than two. The variety of pieces and the extent of the playing field are limitless. Nothing is ever fully in view. The rules of business maneuver and engagement can be changed by any competitor at any time. And for corporations. the competition never ends.

Despite their dramatic differences in complexity. chess and business share two essentials:

— they are systems of competition. and
— the competition involves indirect consequences.

Basic competitive lessons are more readily understood in chess. Yet they are paralleled in business.

Strategy Over Tactics: The Japanese Gambit

In chess. a good strategist excels over a good tactician. Good individual moves do not add up to good play. The strategist sees and manipulates patterns and positions. Certain offensive and defensive chess sequences have become so familiar and predictable that knowledge of appropriate responses is mandatory for skilled players. Truly innovative responses to the Queen's gambit or the Sicilian defense have a high probability of failure. no matter how well played.

In business the Japanese gambit has become equally familiar. Take a complex product. but one for which the technology is fairly mature. Begin with simple goods catering principally to domestic or Southeast Asian markets. Build very high volumes to drive costs down and get quality up. Then attack the North American market at the lower end with these low costs. Roll up the market as domestic manufacturers execute a segment retreat to higher-margined. lower-volume specialties suitable only for their market. Add high-end complexity only at superior total volume and cost. Finally. mop up Europe with the full line of higher-quality. lower-cost goods.

The Japanese gambit became a familiar pattern in transistor radios and televisions. The pattern is being repeated in autos. machine tools. trucks. farm implements. engines. and forklifts. Yet some domestic competitors appear to be executing segment retreats as if the gambit had not been recognized. The best defense against the gambit is strategic offense. not tactical retreat. One response is confrontation in as many markets as possible to choke the volume growth that facilitates the roll-up. A second is to form alliances of complementary strengths. In the automobile industry, Ford seems to be doing the former with the world car and GM the latter with Toyota in California. Both may be more strategic responses to the Japanese gambit than segment retreat.

Competitive Evolution

The strategic value of chess pieces changes during the course of the game. On a crowded early-game board the knight, although limited in range to three squares, is essential because it can skip over pieces to drop behind enemy lines. At that time, the rook, able to strike anywhere on the board along its rank and file, is virtually useless because it has no ability to exercise its inherent range. Later in the game, their utilities are reversed. Pieces fall in battle, enhancing the rook's ability to strike while placing much action out of the knight's range.

The strategic value of various economic competitive advantages also changes as a business evolves. Early on, production experience and cost position may be essential. Later, as markets mature and basic needs are satisfied, differentiated segments can emerge. Marketing may become more important than production. Henry Ford's unassailable cost position with the volume-based Model T strategy was devastated by General Motors during the 1920s with a marketing-oriented strategy of model variation and style change.

Win, Lose or Draw

Many chess games do not result in a win or a loss, but rather in a stalemated draw. This occurs when neither side has sufficient forces left to checkmate the opposition. Inexperienced players frequently fail to recognize these conditions. They continue pointlessly to maneuver. Masters foresee an impending stalemate and either change strategy or agree to stop play.

Impending stalemate can be recognized in business. If a product is relatively undifferentiated and costs are essential, and if the market is large, with many capacity units at maximum scale and with comparable technology, then the industry will have a significant flat section on its supply curve. Stalemate threatens when growth slows. Portions of the forest products, chemicals, metals, and financial-services industries failed to recognize impending stalemate. They continued to invest in unchanged strategies long after such investment was pointless.

STRATEGY AND COMPLEXITY

Direct lessons from chess are applicable in business. More applicable still is the intellectual process by which the game can be mastered. This process describes how an organization plays the business game — and how its performance can be improved.

Many attempts to understand chess have been made by computer scientists and artificial-intelligence researchers. If a computer could play chess well, then perhaps it was intelligent. All early attempts failed miserably.

The problem is permutations. Given rules and positions, it is easy to calculate all possible next moves and select the best. There are always fewer than 1,024 possibilities. It is more than just difficult to do this for the move after next, for that requires calculating all possible responses to the position that results from the best of all possible opposing reactions to one's own best of all possible moves — where the next move may not be clearly best until later, and where the opponent may not choose its best move!

Nevertheless, there are computer progams that play passable chess. Such programs depend on two devices to circumvent the multiplying combinational permutations required to devise a winning strategy. These devices are equally useful in the formulation of business strategies.

Pattern Recognition

One device is recognizing patterns. This is the art of making useful abstractions. Rather than analyzing square by square and piece by piece, a pattern of play becomes the unit of analysis. Strategies rather than tactics are evaluated.

Failure to seek patterns in business results in either analysis paralysis or firefighting. The forest becomes lost in the trees. Patterns filter extraneous information, reduce complexity, focus on the essential. Only key patterns of competitive behavior are evaluated. Chess gambits are patterns of moves. Stalemate is a pattern of low growth in a large commodity market with multiple units of similar cost and capacity and high operating leverage.

Rules of Thumb

The other device is guidelines: rules of thumb. Rules of thumb are guiding principles that, while never strictly true because oversimplified, point reliably to the probable direction of action. Experience builds rules by remembered results of trial and error.

All good chess programs depend on rules of thumb to simplify calculations. Some moves are not examined, because they almost never pay. Very possibly, the brilliantly innovative winning move will be excluded by such a process. But the frustrations of novice players matched against the simplest chess programs demonstrate their power.

Most of the major ideas in business strategy are conceptual rules of thumb about economic competition. The experience curve, the growth-share matrix, average costing, and the environments matrix are all rules of thumb. They help to pattern competitive behavior. They point toward the relevant and away from the extraneous, and suggest probable courses of action. They simplify, but cannot substitute for, the thought process.

Learning

In chess, the ability to learn and adapt still gives grand masters an edge over computer rules. Rules are fixed. Even modern "learning" programs merely readjust the probabilities attached to a predetermined set of rules as their experience accumulates. They cannot rewrite their basic logic.

Most organizations behave like such chess programs. They become bureaucratized. Their decision processes become fixed and difficult to revise.

A learning organization can both refine its pattern recognition and revise its rules of thumb as business competition evolves. The failure of most major businesses does not involve bad initial play; at one time they grew and prospered. Rather it is a failure to learn and adapt as the business changes.

The effort to create successful chess programs parallels the effort of top management to form successful business organizations. Strategic success requires:

- Appropriate pattern recognition. The organization must seek out all relevant information but not be overwhelmed by trivial detail.
- Appropriate rules of thumb. Decision rules must reflect competitive reality at several levels of complexity. Too simple, and decisions will be erroneous; too elaborate, and they will be made late or never.
- Learning. No intelligence networks, reporting systems, filters, or decision rules can be appropriate always, everywhere. Learning when rules do not apply, and when exceptions justify new ones, is the essence of adaptive strategy.

Business is infinitely more complex than chess. It is played continuously, in earnest, for real and significant consequences. Good strategies, flexible responses, and recognition of end games are valuable lessons. Pattern recognition, rules of thumb, and learning are needed to cope with the infinite diversity and indirect consequences of business decisions. The grand masters of business may never truly master it. But they do win consistently over the competition.

Rudyard L. Istvan

Rudyard Istvan is a vice president in the Chicago office of The Boston Consulting Group.

THE WALL STREET JOURNAL MONDAY, JUNE 15, 1998

The Myth of Market Share

Manager's Journal

By Richard Miniter

Jubilant General Motors executives recently announced their May sales figures, showing that GM's share had surged to more than 32% of the U.S. car market. A few months earlier, they were deciding how to deal with a crisis: Domestic market share had fallen to 28.6% in February. Though GM's market share had been sinking since 1992, the executives had assured Wall Street that it wouldn't fall below 30%. It had, and it continued to fall in March.

The GM executives, who promised for years that they wouldn't trade profits for market share, now recanted. They began offering 1.9% financing, fat rebates and thinly profitable lease rates. GM sold thousands of cars at near cost to rental chains—all in the name of market share.

Under pressure from unions and dealers, which want to keep factories humming and cars rolling off the lots, GM is repeating a strategy that cost it tens of millions of dollars in the 1980s and failed to stop the fall of its market share (then 40%). As a result, GM's earnings per share are significantly below those of other auto makers. Last year Ford, with a market share of about 25%, earned $978 more in profit than GM on each car sold.

Flawed Theories

Nearly every public corporation is mesmerized by market share. Keeping it, growing it, justifying it—try to talk to a senior executive without the subject coming up. This obsession, based on flawed theories, drives down profits and pushes businesses into costly moves like GM's. Except in very rare cases, managers should ignore market share. This is heresy to high-powered consultants and business school professors who champion the cult of market share. They make three arguments, all of which are wrong.

The first is "leadership." The theory is that the dominant firm can add features to its product or reduce its marginal profit to control the price. This forgets that consumers, not producers, decide what features are important and how much they are worth. Trying to use market power to tell customers what they need or how much they should pay ultimately topples "market leaders." IBM learned this lesson the hard way in the late 1980s, when it lost ground to Compaq and never regained its pre-eminent position in the personal computer market.

True, market leaders can set prices by reducing marginal profits, thereby increasing sales. But why should they? If the percentage reduction in price per unit is larger than the percentage increase in market share, the firm loses money. This is usually what happens. Seven-Up proved this point with its massive growth in the late 1970s, largely due to discounting. It soon learned that profits didn't follow from leadership and was eventually acquired by Cadbury Schweppes.

A study of more than 3,000 companies finds that more than 70% of the time, the firm with the biggest share of the market doesn't have the highest rate of return.

Cutting margins to boost sales also invites competitors to match the new low price. If the competition's cost structure is the same or lower, then a pointless price war follows. This spiral periodically engulfs the airline industry, as it is doing to GM and other auto makers now. "It's a game of chicken," one GM executive told this newspaper.

The second argument is simple and widely believed: As market share increases, so do profits. This view overlooks comparative advantage and price. There are always markets that are cheaper for a niche player to serve than a mass producer. Why go after the next 10% of market share when the necessary price reductions will bring demands for discounts from the 30% you already serve? And among that 30%, there are almost always customers who cost more than they are worth. In both cases, attempts to maximize market share mean lower profits.

This is a lesson AT&T has yet to learn. The long-distance giant literally buys market share by mailing millions of checks that can be cashed only when the recipient agrees to switch. Of course, MCI and Sprint do the same thing. Savvy customers now switch only long enough to cash the check, or they play the companies off each other, often bidding up the bribe to $100. There are no reliable estimates of how much this costs the phone giants, but market share is the misguided motivation.

The third argument for market share is based on economies of scale: The more widgets you make, the cheaper they are per unit. But economies of scale don't come automatically from bigger market share or greater volume. Doughnuts are cheaper by the dozen only when bakers use the knowledge gained from increased production to shrink costs. Most firms squander executive energy on market-share growth instead of exploiting the economies of scale that are supposed to come with it. "They think of market share as an end in itself. It's not," says Jack High, acting dean of the George Mason University Business School in Fairfax, Va.

To be sure, there are a few situations in which a market-share strategy makes sense. The most persuasive case for such a strategy involves what economists call a "network externality"—when the benefits or price of a product are less important to users than its compatibility with other people and other products. Word-processing software is a case in point. WordPerfect was the leader in market share in the early 1980s, but Microsoft Word was gradually beating smaller rivals like Wordstar and XyWrite. Then Microsoft made Word compatible with its spreadsheet, database and other programs, giving the word processor an advantage and building market share. As Word's share grew, even dedicated WordPerfect users switched so that they could be compatible with the greatest number of users and programs.

In the vast majority of businesses, however, the idea of network externalities doesn't apply. Price, features and other factors matter a lot more than compatibility.

The appeal of market share is partly psychological. It's a lot easier for an executive to win internal support for trimming margins than for making more radical changes to improve a product. Given internal resistance to change, it is often easier to buy market share than to enhance profits. According to economist Tom Rostici of George Mason, executives "feel embarrassed pounding their chest and saying that their company is the most profitable. It's nicer to say that they have the most market share—especially if they aren't that profitable."

Statistical evidence that pursuing market share is folly comes from Donald V. Potter, president of Windermere Associates, a Moraga, Calif.-based consulting firm. Mr. Potter's firm has conducted a groundbreaking study of more than 3,000 public companies. It found that more than 70% of the time, the firm with the biggest share of the market doesn't have the highest rate of return.

Smaller Rivals

The Windermere study examined some 240 industries that have at least five competitors with annual gross sales of more than $50 million apiece. Among the top four firms in each of 240 industries, the market-share leader led the industry in pretax returns on assets only 29% of the time—only 4% better than random chance. In highly concentrated industries, market leaders did slightly better (they had the best industry returns 38% of the time) and in high-growth industries they did slightly worse (only 27% had the best industry returns). Though the study tracked profits in terms of return on assets, the results would be virtually identical if profits were measured in terms of return on equity.

Mr. Potter points to an array of industry leaders bested by their smaller rivals. Steelcase, the nation's largest office-furniture maker, earns about half the rate of return of Hon Industries, which is only one-third its size. Wal-Mart earns lower rates of return than the Family Dollar and Dollar General chains. Merck may be the market leader in pharmaceuticals, but Pfizer is the profit leader.

Profit leaders, even ones with the largest market share, do two things that other big firms don't: They continually trim costs, and they aggressively weed out customers who generate low returns. They ignore market share because they know that size doesn't matter—profits do.

Mr. Miniter is an Alexandria, Va., writer.

WHARTON REPROGRAPHICS